On Quantifying Quality of Care

Machteld Varewyck

Proefschrift voorgedragen tot het behalen van de graad van
Doctor in de Statistische Data-Analyse
Academiejaar 2015-2016

Promotoren:
Prof. dr. Els Goetghebeur
Prof. dr. Stijn Vansteelandt

Vakgroep Toegepaste Wiskunde, Informatica en Statistiek
Faculteit Wetenschappen, Universiteit Gent
Krijgslaan 281, B-9000 Gent

Research funded by a Ph.D. grant of the Agency for Innovation by Science and Technology (IWT)

Cover art by Preben Bonte and Maarten Varewyck

Contents

Dankwoord

Tijdens mijn doctoraat heb ik het voorrecht gehad om niet één, maar wel twee promotoren te hebben die tot de top behoren in hun vakgebied. Dat de lat daardoor hoog werd gelegd, daar ben ik nu oprecht dankbaar voor.

Els, je hebt een belangrijke rol gespeeld in het sturen van mijn jonge loopbaan. Je wist me te motiveren voor maatschappelijk belangrijke thema's tijdens mijn bachelor- en masterthesis. Tijdens dit doctoraat bracht je opnieuw boeiende onderzoeksvragen aan en kon ik op je steun en gedrevenheid rekenen bij het beantwoorden ervan. Bedankt voor de vele uren die je in mij en in dit werk hebt geïnvesteerd.

Stijn, bedankt voor je eindeloze geduld bij het uitwerken van kleine en grote problemen. Als ik door het kluwen van statistische formules soms de uitweg niet meer zag, kon ik altijd bij jou terecht. Het vertrouwen dat je toonde in een goede afloop heeft me doen volhouden tot de eindmeet.

Marie, thank you for providing me access to the Riksstroke data and to initiate me into this brilliant register. Although the answers to our research questions sometimes yielded surprising results, you never doubted my statistical knowledge but instead helped explaining the underlying mechanisms. Tack!

I would also like to thank all members of the examination committee: Marie Eriksson, Els Goetghebeur, Tom Loeys, Sharon-Lise Normand, Rosanna Overholser, Olivier Thas, Herman Van Oyen and Stijn Vansteelandt. You made me realize how easily the bigger picture fades into the background when working on one topic for four years. I highly appreciate your constructive and insightful

comments to this thesis.

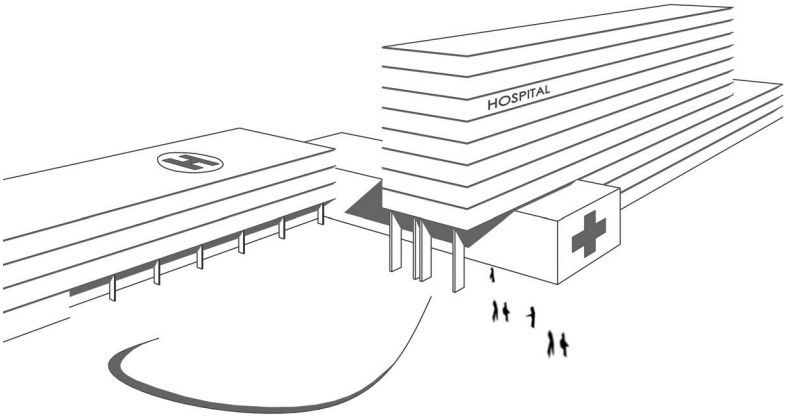
Naast mijn promotoren hebben ook de andere leden van onze onderzoeksgroep me vaak geïnspireerd en gemotiveerd tijdens dit doctoraat. In het bijzonder wil ik Karel bedanken, nog zo een krak in zijn vak! Tijdens onze opleiding hebben we vaak samen gezwoegd op examens en het herwerken van papers. Ik wil je graag bedanken voor je hulp bij het doorgronden van statistische mysteries, maar ook voor je luisterend oor en je schouderklopjes. Onze wegen gaan nu een andere richting uit, maar ik wens je veel succes en denk maar niet dat je nu aan mij zal ontkomen! Bart VR, het is altijd fijn om te kunnen terugvallen op een collega met zoveel ervaring. Door onze gemeenschappelijke Zweedse uitdaging hadden we vaak boeiende discussies. Dankzij de aanwezigheid van Bart, Johan, Jozefien, Karel en Sjouke hebben de vele conferenties me niet alleen op statistisch vlak iets bijgebracht. Ook al heb ik veel bureaugenoten versleten, ze mochten er stuk voor stuk wezen. Bedankt Jan, Lizzy, Peter, Koen, Jens en Pieter om de soms lange dagen iets korter te maken en de soms grote frustraties iets kleiner te maken. Jan, je liet me vanaf de eerste dag thuis voelen op ons bureau en al was je parcours soms onvoorspelbaar, ik ben supertrots op jou. Daarnaast wil ik ook alle andere TWIST-leden bedanken. Catherine, jij bent de grootste 'welbevinden op het werk'-heldin die ik ken. De spelletjesavonden, weekends en barbecues zorgen ervoor dat iedereen zich hier thuis voelt. Als ik ooit een wedding-planner nodig heb, weet ik je te vinden! Virginie, bedankt voor je onuitputtelijke energie, je talloze mailtjes en natuurlijk een dikke dankjewel voor je vastberadenheid om me van straat te helpen. Charlotte, onze middagpauzes waren altijd veel te snel om, net als ons tweejaarlijks stoffenjacht-avontuur. Herman, JAMES is echt supercool, bedankt om hem draaiende te krijgen op mijn computer. Bedankt ook sportieve vrienden voor al dat badmintongeweld, wat zweet mocht zeker niet ontbreken in een opleiding als deze.

Ook buiten het werk stonden er vrienden klaar voor wat ontspanning nu en dan. Bedankt Elke, Frauke, Jolijn, Nathalie en Sara om geduldig naar een geschikt moment te zoeken zodat iedereen mee kon op vrijgezellen, naar de kerstmarkt of een Lokers etentje. Lotte, bedankt voor de warme omhelzingen als het even moeilijk ging, maar ook voor de spontane afspraakjes die altijd een groot feest werden. Stijn, bedankt voor de vele jog-zondagen zodat we met een fris hoofd

aan de nieuwe week konden beginnen. De muzikanten van het GUHO, bedankt om het stof van mijn dwarsfluit te blazen en me te laten meegenieten van jullie prachtige orkest en de onvergetelijke concerten.

Liefste mama en papa, duizendmaal dank om bij te springen als het onverwachts druk, moeilijk of spannend werd. Dat ik altijd op jullie kan rekenen is zodanig vanzelfsprekend geworden dat het veel te weinig wordt gezegd, dankjewel! Bedankt om me zoveel te leren naast mijn studies en me te steunen in alles wat ik doe. Bedankt Matthias en Ellen, Maarten en Evi voor de gezellige familiemomenten en jullie interesse in mijn onderzoek. Bedankt ook aan jullie kroost die me elke keer weer doet beseffen wat echt belangrijk is. Bedankt lieve Preben, om mij geduldig te laten uitrazen na een werkdag, om me plat te knuffelen als ik te veel pieker en om een unieke covertekening te maken. Ik ben helemaal klaar voor een nieuw avontuur, samen met jou!

Machteld Varewyck
December 2015



Glossary of Acronyms

CC	Complete cases
CI	Confidence interval
DSR	Directly standardized risk
FE	Fixed effects
ISR	Indirectly standardized risk
MAR	Missing at random
MCAR	Missing completely at random
ME	Mixed effects
MI	Multiple imputed
PS	Propensity score
RE	Random effects
SMR	Standardized mortality risk

CHAPTER 1

Introduction

1.1 A Global Need for Quantifying Performance

Improving the quality of health care has been captivating people for ages, just as it did Florence Nightingale, a great pioneer in statistics:

"In 1855, Florence Nightingale tabulated the death rates from leg amputation of English soldiers who participated in the Crimean War. She observed that after accounting for the level of amputation above or below the knee, soldiers operated on in large hospitals were more likely to die than those operated on in small hospitals. She identified the causes of this unexpected finding as poor sanitation and the rapid spread of infection from patient to patient in large hospitals. She pleaded with English royalty to do something about the sanitary conditions of English field hospitals." (from Keeler et al. (1992))

More than a century later, there continues to be great interest in characterizing hospitals that experience a better or worse outcome than expected and to use this information in some way to improve health care: For example, giving feedback to hospitals can help target interventions such as continuing education for the personnel, government agencies on the other hand can make evidence-

based decisions on public expenditure, while public reporting may guide patients in their hospital choice (Normand et al., 1997). Implications may thus be far-reaching whereby criticism of hospitals may damage staff morale and public confidence or even lead to closure (Black, 2010). For example, in the United Kingdom (UK), reports on poor performance led to the dissolution of the Mid Staffordshire NHS Trust and transferring services to other health trusts [1]. In the United States (US), St. Mary's Medical Center in Florida permanently closed its pediatric cardiothoracic surgery program after exceedingly high mortality rates had been published [2].

Measuring hospital performance is not a new activity, neither is it a local phenomenon. Several initiatives have been set up, nationally (e.g. National Health Service in the UK [3]) and internationally (e.g. the PATH project in Europe [4]), all sharing the goal to improve health care. In its recommendations, the World Health Organization (WHO) European Region stated that by the year 2010 all countries should have a nationwide mechanism for continuous monitoring and development of the quality of care for at least ten major health conditions (World Health Organization and others, 2003). Results on hospital performance have been published since the mid-1980's in the United States and since 1999 in the United Kingdom. In the last 10 years, many other countries have introduced various forms of hospital performance measurement. In Belgium, the Flemish government has the authority for health care and public welfare. Recently a website was launched to compare the quality of Flemish hospitals [5], where in a first phase most reports are on breast cancer treatment. Not only the quality of hospitals is increasingly monitored, also other public institutions such as nursing homes, universities and schools can learn from each other (Leckie and Goldstein, 2009; Spiegelhalter, 2005a). The social relevance and broad applicability only emphasize the importance of accurate data analysis that allows for making fair comparisons. Although we will focus on hospital performance, this specific context is exemplary rather than restrictive.

[1] http://www.bbc.com/news/uk-england-stoke-staffordshire-23508096
[2] http://edition.cnn.com/2015/08/17/health/st-marys-medical-center-investigation
[3] http://www.nhs.uk
[4] http://www.pathqualityproject.eu
[5] http://www.zorgkwaliteit.be

Quantifying a hospital's performance by its crude observed mortality rate is naive and may be highly misleading. Indeed, a hospital may show a high mortality rate because it mostly treats severely ill patients and not because it has poor quality of care. In Section 1.3 we briefly explain the standard statistical methods to handle this and other challenges when quantifying performance. Some of their drawbacks motivate the developments we present in the following chapters. Before discussing analysis methods, we introduce Riksstroke, the Swedish register for stroke patients, which will be analyzed in the following chapters and illustrate the discussed methods.

1.2 Riksstroke, the Swedish Register for Stroke

Patient attributes may partly explain differences in hospital outcomes. Not accounting for these differences between patients at hospital admission would result in unfair comparisons, which is clearly not a good idea. So, for reliable assessment of hospital performance, accurate registration of patient data is the basis (Brookhart et al., 2010). Efforts towards more systematic and accurate registration have been facilitated by digitization (Iezzoni, 1997). Electronic health records for example allow to share patient data across different health care settings. Sweden has been a pioneer with a strong tradition in high-quality registers. One of these is Riksstroke, the national register for stroke patients that will be analyzed throughout this thesis.

Riksstroke, the Swedish quality register for stroke care, aims to monitor and improve hospital performance and ultimately to ensure the best possible care for stroke patients (Asplund et al., 2011). It is one of the world's largest stroke registers. It is estimated to cover between 80% and 90% of the total stroke population with considerable variation in coverage between hospitals and over the years since 2001. A detailed analysis of Riksstroke data quality showed that data entry errors were marginal: A 95% consistency rate was noted upon comparing the diagnosis stated in Riksstroke with the medical chart.

We consider different subsets of the data across the chapters, either because the data were updated with the most recent records or new patient characteristics

had only recently been measured. In general, we restrict our analysis to the first registered stroke for adult patients (≥ 18 years), treated in one of the 90 Swedish hospitals between 2001 and 2012. We consider patients diagnosed with ischemic stroke (ICD-10 I63), intracerebral haemorrhage (ICD-10 I61) or unspecified acute cerebrovascular event (ICD-10 I64). This dataset contains 249 414 patients. Although the sample size is large, the number of registered patients per hospital can be problematically small e.g. for year-specific analysis. Using personal identification numbers, records in Riksstroke are linked with the register from Statistics Sweden so that a wide range of information is available on the patient's background (e.g. age, yearly income, education level), medical treatments during the hospital stay (e.g. stroke unit care, thrombolysis) and patient's follow up (e.g. 30-day mortality, living conditions at 3 and 12 months). Generally speaking, the register shows limited differences in patient mix across hospitals (see Chapter 2), because no strong regional differences in patient characteristics are present and stroke is an acute disease so that patients are mostly treated in the nearest hospital.

Analytical reports on process and outcome quality indicators are accessible for the public on the Riksstroke website: `http://www.riksstroke.org/eng/`. To facilitate data interpretation, background information on the hospitals is presented, such as patient characteristics and coverage of the total stroke population. Key quality indicators in Riksstroke are also presented in a report on the quality of Swedish healthcare produced annually by the National Board of Health and Welfare and the Swedish Association of Local Authorities and Regions. Participating hospitals have access to reports presenting their own data in more detail and in comparison with national data. A range of reports are thus provided, each adapted to the stakeholders' interest and background knowledge.

1.3 Statistical Issues when Quantifying Performance

The assessment of hospital performance concerns many different stakeholders such as hospitals and clinicians, government and health insurers, and last but not least, the patients and their support group. They may all have different moti-

vations for learning about hospital performance such as reducing the workload, identifying hospitals that need extra investments or improving patient satisfaction. To perform a proper statistical analysis and come to relevant reporting, close collaboration with the target audience is thus needed. In our case, frequent consultations with Marie Eriksson, a member of the Riksstroke statistics team, and visits from/to Sweden were arranged to gain insight in the Swedish register Riksstroke.

1.3.1 Deciding on the indicator of interest

Hospital performance cannot be captured in one number, so a range of indicators are used that are thought to be related to the given care level. Untimely death is often used in this context, because it is easily measured, of undisputed importance and encountered across hospitals (Lilford and Pronovost, 2010). However, in settings where all hospitals are expected to have similarly high or low death risks it is certainly not the ideal indicator. For example when patients go to the hospital for end-of-life care, everyone is expected to die within a short time span so that quality of life may be a better indicator of good hospital performance. Also when (almost) no deaths are observed within a relatively short time frame, e.g. for 1-day surgical procedures or chronic diseases, patient satisfaction is probably a more relevant outcome measure. Outcome indicators are frequently criticized as they do not directly point to how, if so, the given care could be improved (Freeman, 2002). Still, hospitals can utilize them for internal evaluation and start searching for reasons that may explain differences between their and other centers' outcome indicators and what they can learn from it.

Besides outcome indicators, which assess whether the given care is effective, process indicators assess whether patients get the care they need as indicated in guidelines, e.g. 'Is blood pressure measured daily?' (Campbell et al., 2000). These process indicators are more direct measures of quality of care, although it is not always obvious which guidelines have to be met, e.g. Is it necessary to measure blood pressure daily? For further discussion on when to best use which type of indicator, we refer to Lilford et al. (2004), Mant (2001) and Campbell et al. (2000). In either case, it is important to focus on one department or disease of

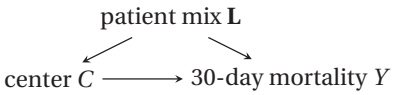
the hospital at a time, rather than aiming to evaluate the hospital as a whole. This because differences in the quality of care within hospitals are often much larger than differences between hospitals (Jha et al., 2005).

Of course, multiple outcome indicators or a combination of outcome and process indicators will provide better insight in the quality of care. Anyhow, some selection is required as too many indicators may complicate the interpretation and increase the maintenance cost (Freeman, 2002). We will focus on 30-day mortality following a specific disease (e.g. stroke), denoted by Y (1 = death, 0 = survived), but the methods we will discuss are also applicable for other outcome indicators.

> **Example 1.1**
> On the Flemish website for hospital comparisons (www.zorgkwaliteit.be) concerning breast cancer treatments, both process indicators (e.g. percentage breast saving surgery) and outcome indicators (e.g. 5-year survival probability) are reported per hospital, see also Figure 1.1.

1.3.2 Defining the causal inference framework

In general, a comparison of crude mortality risks between hospitals can only be fair if their patient mix (e.g. age, initial disease severity) is very similar. Only then can higher observed mortality risks effectively be attributed to worse quality of care. In practice, however, this is very exceptional, so that other statistical measures are needed that do account for baseline differences in patient mix among hospitals. The causal effect of a hospital on the outcome indicator will be our target focus: It expresses the differences in mortality risk that cannot be explained by differences in patient mix, but by differences in hospitals. This effect is represented by the arrow from the hospital center C to the outcome Y below:



GEOBSERVEERDE VIJFJAARSOVERLEVING GECORRIGEERD VOOR LEEFTIJD EN STADIUM

Hoeveel procent van de patiënten is nog in leven vijf jaar na het vaststellen van borstkanker, rekening houdend met de leeftijd van de patiënt en de uitgebreidheid van de tumor? Het gaat om een schatting. Bij deze indicator tellen alle doodsoorzaken mee (niet alleen kanker). Bekijk deze samen met de andere indicatoren over overleving.

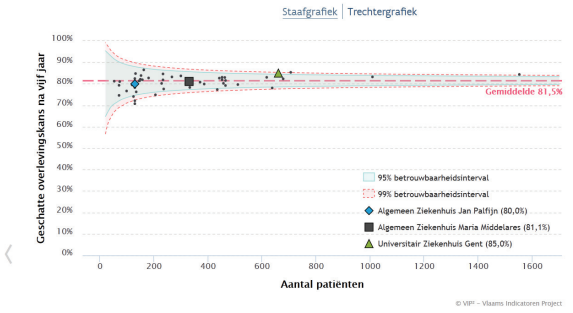
Waarom is deze indicator belangrijk?

De kans dat patiënten vijf jaar na het vaststellen van borstkanker nog leven, hangt in belangrijke mate af van de leeftijd van de patiënt en het stadium van de ziekte (= de uitgebreidheid van de tumor).

Het is belangrijk om deze twee factoren in rekening te brengen bij de vergelijking van overlevingscijfers tussen ziekenhuizen. Indien het ene ziekenhuis gemiddeld gezien oudere patiënten en/of meer gevorderde borstkankers behandelt dan een ander ziekenhuis, dan is het te verwachten dat de overlevingskans in het eerste ziekenhuis lager ligt dan in het tweede. Zonder deze verschillen tussen ziekenhuizen in rekening te brengen, kan geen eerlijke vergelijking tussen de ziekenhuizen worden uitgevoerd.

Wel dient opgemerkt te worden dat het onmogelijk is om te corrigeren voor alle factoren die van belang kunnen zijn in de overleving, aangezien deze informatie simpelweg niet allemaal voorhanden is (bv. de sociale status van de patiënt). Verder wordt in deze indicator rekening gehouden met alle doodsoorzaken en niet enkel met deze ten gevolge van de borstkanker zelf. Daarom is het ook belangrijk om naar de indicator over de relatieve overleving te kijken.

Voor deze indicator is er geen streefwaarde.



LEGENDE

- Dit is een zogenaamde **trechtergrafiek**.
- Hoe hoger een ziekenhuis in deze grafiek ligt, hoe hoger zijn resultaat (**de verticale as**).
- Hoe meer naar rechts het ziekenhuis ligt, hoe meer patiënten waarvoor de indicator gemeten is, en hoe betrouwbaarder het resultaat (**de horizontale as**).
- **Het 95% betrouwbaarheidsinterval**: elk ziekenhuis dat binnen deze trechter ligt, heeft geen afwijkend resultaat in vergelijking met alle andere ziekenhuizen.
- **De rode stippellijn**: het gemiddelde van alle resultaten.
- **De zwarte stippen**: de resultaten van alle andere ziekenhuizen.
- **Het blauw gearceerde veld**: dit is de streefwaarde voor deze indicator (hoe wordt de streefwaarde gekozen?)

Source: www.zorgkwaliteit.be

Figure 1.1: Hospital quality of care assessment in Flanders: 5-year survival probability following breast cancer, controlled for age and disease stage.

In order to capture the causal center effect it is important to control for the differences in patient mix **L** (DeLong et al., 1997; Austin et al., 2003). In a causal inference setting, this is better known as controlling for confounding.

In fact, many possible confounders may exist, some of the most common are age, gender, initial disease status, smoking status, whether the patient has diabetes or any other chronic disease. However, only those patient characteristics (i.e. confounders) that were actually measured can be incorporated in the outcome regression model. To estimate the causal center effect we will therefore assume that the measured patient characteristics are sufficient to adjust for confounding of the center-outcome effect (Hernán and Robins, 2006b). This assumption of 'no unmeasured confounding given the measured variables' states that patient outcomes are conditionally exchangeable within levels of **L**. More specifically, it is assumed that for patients with the same baseline characteristics, the mortality risk of those treated under the care level of center A, had they been treated under the care level of center B, would have been the same as the observed mortality risk in center B. Expert knowledge is needed to identify the potential confounders already at the design stage of the study, although full registration may still be limited, not only by lack of information but also by budget and workload constraints. This assumption may be violated when important confounders are not measured. For example, patient's socioeconomic status is rarely measured although it has recently been shown to affect the mortality risk of stroke patients in Sweden (Lindmark et al., 2014) and it is not unlikely to also influence at which hospital the patient is actually treated as wealth may differ between geographical regions.

Two other assumptions for causal inference are the 'stable unit treatment value assumption' (SUTVA) and the positivity assumption. Let $Y(c)$ indicate the potential outcome for a given patient if treated at the care level of center c. Then, the SUTVA states that a patient's potential outcome $Y(c)$ at care level c does not depend on other patients' given care level (Hernán and Robins, 2006b). For Riksstroke the interaction between patients is minimal: We restrict our analysis to the first registered stroke for patients. If not, the potential outcome for a single patient with a second stroke may depend on the given care level for his/her first stroke, e.g. if repeated thrombolysis treatments are not recommended. Another

possible source for violation of the SUTVA assumption are infectious diseases, e.g. if a patient at a given center has the flu, other patients who would be treated at that center may get infected, thereby affecting their potential outcome. The positivity assumption states that for each value of the covariates **L** in the population and for any center c, the probability that some patients with **L** experienced the care level at center c is positive (Hernán and Robins, 2006b). This assumption can be assessed empirically by investigating how patient-mix differs across centers. For Riksstroke, it will be shown (see Chapter 2) that differences in patient-mix across centers are limited, in favor of the positivity assumption. If on the other hand, patient mix would differ across centers, e.g. people in northern Sweden are older than in the southern regions, it is still likely that a patient from northern Sweden is treated at a hospital in the south, because he/she works there or is on vacation while having an acute stroke. In other settings or for other diseases this assumption may be violated e.g. when some centers are not permitted to treat patients who need advanced care.

1.3.3 Two summary measures for performance

Although the causal hospital effect is our main focus, it is not an intuitive measure of e.g. the number of additional deaths compared to the national average. In the field of health services research, these hospital effects will therefore be communicated through directly and indirectly standardized risks, which are the most commonly used summary measures for performance (Keiding and Clayton, 2014).

To illustrate the two standardization techniques, we introduce a simple example in Table 1.1. We consider 3000 patients who were treated in one of 3 hospitals and 30-day mortality was registered as outcome quality indicator. The patient mix across hospitals only differs by patient's baseline disease severity, which is either low or high. It can be seen that hospital 1 and 2 have the same severity-specific mortality risks, respectively 1% for patients entering with low risk and 10% for those with high risk. However, hospital 1 shows a much higher crude mortality risk (9.1%) than hospital 2 (1.9%). This is because hospital 1 treated relatively more patients at high baseline risk than hospital 2, resulting in a larger

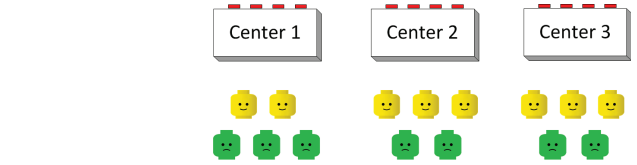
number of observed deaths. So the observed difference in crude mortality risks is not due to a worse care level but due to a different patient mix. In contrast, hospital 2 and 3 have the same patient mix, so pairwise comparisons between their crude mortality risks are fair.

	Baseline disease severity		Total	$\hat{\text{ISR}}$	$\hat{\text{DSR}}$
	Low	High			
Hospital 1				0.93	4.3%
No. patients	100	900	1000		
Deaths	1 (1%)	90 (10%)	91 (9.1%)		
Hospital 2				0.84	4.3%
No. patients	900	100	1000		
Deaths	9 (1%)	10 (10%)	19 (1.9%)		
Hospital 3				1.32	5.7%
No. patients	900	100	1000		
Deaths	18 (2%)	12 (12%)	30 (3.0%)		
Overall					
No. patients	1900	1100	3000		
Deaths	28 (1.5%)	112 (10.2%)	140 (4.7%)		

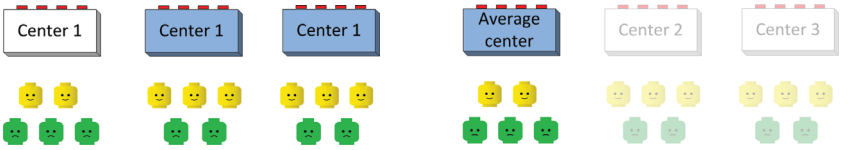
Table 1.1: Toy example comparing the crude, directly and indirectly standardized mortality risks of three hospitals.

Direct standardization (Nicholl et al., 2013; Bos et al., 2005) aims to infer the potential full population risk for each hospital: the risk that would be realized if all patients under study were to experience the care level of that given hospital, irrespective of where they were actually treated. This is illustrated in Figure 1.2b, where the care level of hospital 1 is extrapolated to the patients of the other centers under study, to estimate the directly standardized risk of center 1. For the example in Table 1.1, the directly standardized risk (DSR) for hospital 1 is simply calculated as:

$$\begin{aligned} \hat{\text{DSR}}_1 &= \frac{0.01 \times (100 + 900 + 900) + 0.10 \times (900 + 100 + 100)}{3000} \\ &= 0.01 \times 0.63 + 0.10 \times 0.37 \\ &= 4.3\%, \end{aligned} \tag{1.1}$$



(a) Observed data.



(b) Directly standardized risk of center 1.

(c) Indirectly standardized risk of center 1.

Figure 1.2: Sample of three centers with a different mix of patients having good health condition on admission (yellow smiling face) and poor health condition on admission (green sad face).

and this is the same for hospital 2, reflecting similar care delivered at these two centers. Only hospital 3 has a larger estimated DSR than the overall mortality risk of 4.7%. Direct standardization thus applies for each hospital the same set of weights (resp. 0.63 for the low and 0.37 for the high risk group), but applies the hospital's severity-specific risks. The DSR is therefore insensitive to differences in the hospitals' distribution of baseline disease severity. It is sometimes criticized as not evaluating the hospitals on the patients they actually treated. For example, hospital 1 mostly treated patients at high risk, although for $\widehat{\text{DSR}}_1$ in (1.1) most weight is given to its mortality risk for patients with low baseline disease severity. Formally, the directly standardized risk for center c is denoted by $E\{Y(c)\}$, where $E\{\cdot\}$ refers to the expected value over all patients under study. Given the DSRs, we can easily make pairwise comparisons either among different centers, with past performance or with the overall mortality risk $E(Y)$.

In contrast, indirect standardization (Shahian et al., 2001; Campbell et al., 2012) focuses on what a center achieves for its own patient mix (Figure 1.2c). In general, a risk ratio or risk difference is calculated between the observed and expected risk (e.g. the risk when the average of all observed care levels would apply) for each center. For example, when the indirectly standardized risk (ISR)

is calculated as the ratio of observed and expected mortality risk, we obtain for hospital 1 in Table 1.1:

$$\begin{aligned} \hat{\text{observed}}\ \text{risk}_1 &= \frac{0.01 \times 100 + 0.10 \times 900}{1000} = 9.1\% \\ \hat{\text{expected}}\ \text{risk}_1 &= \frac{(0.01 + 0.01 + 0.02)/3 \times 100 + (0.10 + 0.10 + 0.12)/3 \times 900}{1000} \\ &= 9.7\% \\ \hat{\text{ISR}}_1 &= \frac{9.1}{9.7} = 0.93, \end{aligned} \tag{1.2}$$

stating that the observed mortality risk in hospital 1 is better than the expected mortality risk if its patients were to experience the average of all observed care levels. Similarly, hospital 2 has an ISR smaller than 1, contrary to hospital 3 which has a worse care level than expected. Note that even though hospital 1 and 2 perform equally well for each patient type, their estimated ISRs differ because their patient case-mix differs. Comparing two hospitals based on their ISRs can thus only be fair if their patient distributions are very similar. What indirect standardization actually aims to answer is: "How would the risk in a given center change if its patients were to experience the average risk across all centers?". The observed risk in center c can formally be denoted by $E\{Y|C = c\} = E\{Y(c)|C = c\}$ and the expected risk by $m^{-1}\sum_{c^*=1}^{m} E\{Y(c^*)|C = c\}$, where m denotes the number of hospitals under study.

In summary, direct and indirect standardization are each others mirror-image: For direct standardization, the hospital of interest provides the risks and the total study population provides the weights. For indirect standardization, the population of observed centers provides the risks and the hospital of interest provides the weights. So, the different standardizations may yield different results, because they intend to answer different research questions and pursue different standards. For example, given direct standardization a hospital may show a smaller mortality risk than the overall mortality risk, while this hospital may have an ISR larger than 1, indicating worse performance than the expected mortality risk for its patient mix. This emphasizes the importance of transparency to the end user (Manktelow et al., 2014), otherwise such 'conflicting' results may strengthen the myth that statistics can prove anything.

For both standardization methods, caution is needed when some patient-groups are almost empty or have close to zero events, because then results may be unstable. For example, when hospital 1 in Table 1.1 had 0 instead of 1 observed death for its low risk patients, $\hat{\text{DSR}}_1$ would drop from 4.3% to 3.7% while $\hat{\text{ISR}}_1$ would hardly change (still 0.93). At the same time, the $\hat{\text{DSR}}$ for the other hospitals would not change, while the estimated ISRs would increase: for hospital 2 from 0.84 to 0.97 and for hospital 3 from 1.32 to 1.53.

Example 1.2

- Traditionally indirectly standardized measures are used for national hospital performance evaluations, e.g. in the UK (Clinical Indicators Team, 2015) and the US (Ash et al., 2012). This is indeed the most relevant measure to judge a hospital's performance on its own patient mix or for yearly evaluations, where the latter should ideally come with a careful description of changes in patient mix over time.
- Recently, the usefulness of direct standardization in this context has been recognized (Nicholl et al., 2013), e.g. in Figure 1.1 directly standardized 5-year survival probabilities are reported. It is a valuable measure for example when hospitals vary in approach and one wishes to choose one approach for implementation across all hospitals. The given care level can then be extrapolated to a broader set of patients through direct standardization.

A further discussion of when to best use which standardization is given in Chapters 2 and 3.

1.3.4 Outcome regression modeling to adjust for patient mix

The arithmetic approach we used in the previous section to estimate the DSR and ISR has some practical limitations: Calculations become complicated when the number of patient characteristics is large. Moreover, patient attributes that were measured on a continuous scale (e.g. age) have to be categorized (e.g. 10-year age

groups), reducing their informative value. We have also shown that the calculated standardized risks may become unstable when the observed number of deaths is small for some patient subgroups. In each of these situations outcome regression modeling can help estimate the directly and indirectly standardized risks.

A simple regression model for a patient's outcome Y_i (e.g. patient satisfaction score) given his/her age_i and at which hospital C_i (1 or 2) he/she was treated, is:

$$E(Y_i|\text{age}_i, C_i) = \beta\,\text{age}_i + \psi_1 I(C_i = 1) + \psi_2 I(C_i = 2), \tag{1.3}$$

where $I(C_i = 1)$ is 1 if the i-th patient was treated at hospital 1 and 0 otherwise, similarly for $I(C_i = 2)$. Under the 'no unmeasured confounders' assumption the model parameters (β, ψ_1, ψ_2) can be interpreted as the causal effect of respectively age, the care level of hospital 1 and the care level of hospital 2 on the patient satisfaction. In Chapter 2 and Chapter 3 or in Roalfe et al. (2008) it is explained how the directly and indirectly standardized risks can be calculated based on outcome regression models. For the simple model in (1.3), a possible way to calculate the indirectly standardized risk (as a ratio of observed and expected risk) for hospital 1 is:

$$\hat{\text{ISR}}_1 = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i}{\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\beta\,\text{age}_i + \frac{\psi_1+\psi_2}{2}\right)}, \tag{1.4}$$

where n_1 is the number of registered patients in hospital 1.

The limitations of the arithmetic approach can be overcome by using outcome regression modeling: The outcome model can include whatever type of patient characteristic, such as binary (e.g. smoking status), categorical (e.g. education level) or continuous covariates (e.g. age). Neither is the outcome format limiting, even though we formulated a regression model for a continuous outcome, statistical models exist for binary outcomes (e.g. logistic regression for 30-day mortality, Agresti (2002)) or survival outcomes (e.g. Cox proportional hazards model for time to death, Collett (2015)). For a large number of patient characteristics p and a general number of centers m, the simple outcome regres-

sion model in (1.3) can easily be extended:

$$E(Y_i|\mathbf{L}_i, C_i) = \sum_{j=1}^{p} \beta_j L_{ij} + \sum_{c=1}^{m} \psi_c I(C_i = c), \tag{1.5}$$

where $\mathbf{L}_i = (L_{i1}, \ldots, L_{ip})$ is the vector of confounders for patient i. However, the important challenge is to accurately estimate the model parameters $(\beta_1, \ldots, \beta_p, \psi_1, \ldots, \psi_m)$.

Given a sample of n patients, the model parameters can be estimated for example via maximum likelihood estimation (Neter et al., 1996), which is a widely-used method and implemented in most statistical software programs. When the number of hospitals under study m is large, many hospital effects $\psi_c (c = 1, \ldots, m)$ need to be estimated. Moreover, due to a limited sample size, the number of registered patients may be small for some hospitals so that estimating hospital effects may become impossible or hold unstable and inaccurate results (Peduzzi et al., 1996). Especially the smallest hospitals may suffer from this lack of information. Excluding those hospitals is not an option as smaller hospitals have shown lower care levels in some settings e.g. due to a smaller learning effect for surgical procedures (Silber et al., 2010; Kressner et al., 2009). Currently, the most popular solution is the mixed effects model (Ohlssen et al., 2007a; Mohammed et al., 2012). It assumes fixed effects for the patient characteristics as before, but normal random effects for the hospital, i.e. the hospital effects are assumed to follow a normal distribution: $\psi_c \sim N(\mu, \sigma^2)$. Then, only two parameters (μ, σ) need to be estimated instead of m separate hospital effects $\psi_c (c = 1, \ldots, m)$. To estimate the effect of one specific hospital, information from all other hospitals under study is borrowed, namely via μ and σ. However, this approach may shrink the hospital effect towards the 'average' hospital effect μ (Kalbfleisch and Wolfe, 2013; Ash et al., 2012), especially for the smallest hospitals, as they put less weight in estimating μ and σ. For example, in Figure 1.3 we plot a histogram of hospital effects ψ_c $(c = 1, \ldots, m)$ together with a normal distribution curve having sample mean 0.05 and sample variance 1.19. It is clear that when the hospital effects are drawn from the approximating normal distribution, this results in an accurate estimate for most hospital effects, but the percentage of

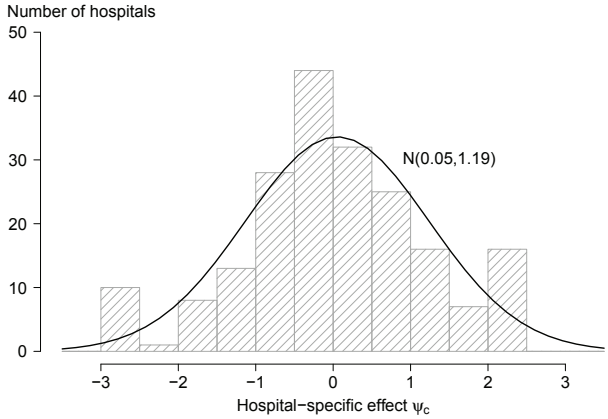


Figure 1.3: The distribution of the hospital-specific effects ψ_c ($c = 1, \ldots, m$) for $m = 200$ hospitals and an overlaying normal distribution function.

hospitals with an effect in the tails will be smaller than the observed percentage. Thus, outlying (good or bad) performance may be masked which is of course undesirable. An alternative method to estimate the hospital effects is via Firth corrected maximum likelihood estimation (Firth, 1993), which does not assume a normal distribution for the hospital effects, but instead a distribution that has heavier tails, allowing for more substantial deviations from the 'average' hospital effect. This method builds upon ordinary maximum likelihood estimation, but is developed to reduce bias on the parameter estimates and has shown to fix convergence problems even when only few events are observed in some patient subgroups. Its use in the context of assessing hospital performance will therefore be investigated in Chapter 2.

Example 1.3

- In the US, the Centers for Medicare and Medicaid Services estimate the expected risk based on a mixed effects model with fixed patient effects and random hospital-specific effects (without including hospital

attributes such as volume) (Ash et al., 2012).

- In the UK, the Health and Social Care Information Centre reports that the expected risk is based on a fixed effects model with patient effects and without hospital-specific effects (Clinical Indicators Team, 2015).

1.3.5 Benchmarking and reporting

Given the estimated standardized risks, we also need to be aware of their precision: If the average mortality risk is 20%, a standardized mortality risk of 25% is in general more worrisome for a hospital with 400 registered patients than for a hospital with 100 registered patients. This source of uncertainty on the estimated standardized mortality risks (SMR) can be expressed through its variance, whereby smaller hospitals will have a larger variance. We can then easily construct a confidence interval for each hospital's SMR. Traditionally 95% confidence intervals are used, which express a range of values (e.g. 22% to 28%) that you can be 95% certain contains the true hospital's SMR, which is unknown. So, if we would be able to take 100 data samples, then on average 95 of the constructed confidence intervals will contain the true SMR.

Funnel plots are used in an increasing number of applications to provide a graphical representation of hospital performance (Spiegelhalter, 2005a; Campbell et al., 2012). Examples are given in Figure 1.1 and Figure 1.4, where the hospital performance measure (e.g. in Figure 1.1 estimated 5-year survival probability) is plotted against a measure for its precision (e.g. the number of registered patients), a horizontal line is drawn at the average (e.g. 81.5%) and thresholds indicate for which hospitals the performance measure is significantly different from that average. Over-dispersion may dilute this plot, i.e. when the variability in hospital performance measures is so large that it cannot be attributed to chance and a few divergent institutions (Spiegelhalter, 2005b). Then the majority of hospitals lies outside the threshold values, which makes the labeling useless. Several methods exist to temper over-dispersion, from ad-hoc methods that cluster the hospitals in more homogeneous groups, to more advanced methods that estimate an over-dispersion factor which expands the control limits

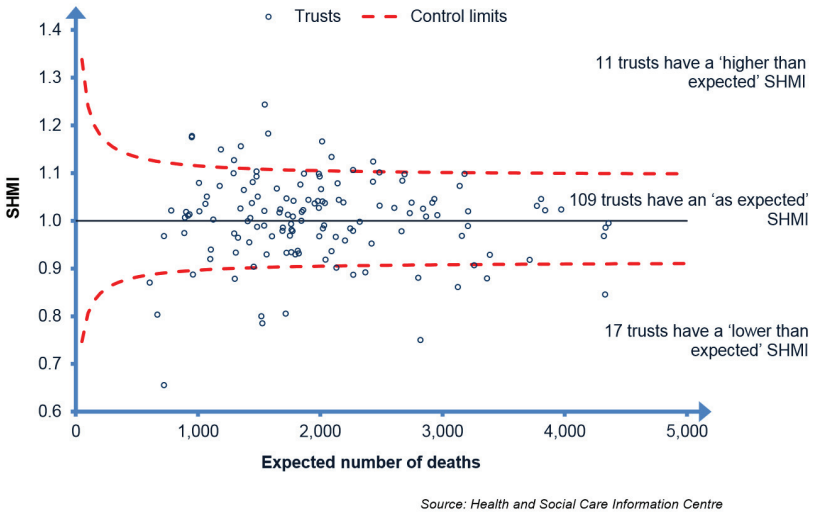


Figure 1.4: Summary Hospital-level Mortality Indicator (SHMI) funnel plot for the period January 2014 until December 2014.

(Spiegelhalter, 2005b).

Alternatively, a range of clinical equivalence can be defined (Normand et al., 1997), for example in Figure 1.5 the clinical equivalence limits are set at 0.8 and 1.2 times the average mortality risk $\hat{E}(Y)$. Then, for hospitals with an estimated SMR outside the clinical equivalence zone, we aim to express how certain we are about the estimated SMR and whether its true SMR is expected to exceed the clinical equivalence limits. Therefore, a hospital is labeled as having high mortality risk if (the lower limit of) its 50% confidence interval exceeds the upper clinical boundary, similarly for low mortality risks. Taking into account the magnitude of effect is especially important because e.g. large hospitals tend to have very narrow confidence intervals which are virtually guaranteed to exclude the average risk even though such differences may not be clinically relevant. When taking into account a clinical equivalence range, the level of confidence should accordingly be decreased (e.g. from 95% to 50%), otherwise hardly any hospital would be labeled as having outlying performance, so that the labeling

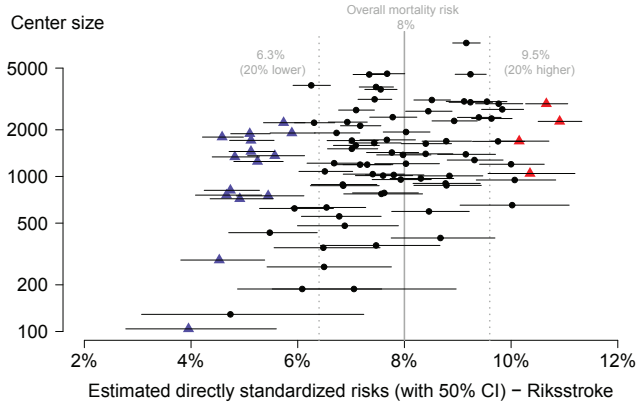


Figure 1.5: A hospital is labeled as having low (blue) or high (red) mortality risk if the 50% *confidence interval (CI) for its directly standardized risk respectively exceeds the clinically relevant boundaries* 20% *lower or higher than the overall mortality risk of* 8%.

would again become useless.

The chosen values for the statistical and clinical relevance determine how much priority is given to avoiding Type I and Type II errors. A Type I error is made if a hospital with acceptable performance is labeled as having low/high mortality risk. On the other hand, Type II errors occur when hospitals with low/high mortality risk are not detected as such. When reports are openly published, Type I errors may be more harmful than Type II errors, because wrongly labeling a center as having high mortality risk may lead to unjust negative advertising. A smaller percentage of Type I errors may then be obtained by either constructing wider confidence intervals or widening the clinically meaningful interval. However, these actions will increase the percentage of Type II errors. If results are used for internal evaluation in hospitals, the stakeholders can decide to set the benchmarking values so that the percentage of Type II errors is small and hospitals with deviating performance have a high probability to be detected as such.

Example 1.4

- In the US, comparative performance information in the form of 'report cards' (e.g. `http://www.valleyhealthlink.com/usnews`) has been published for the public for over a decade (Normand and Shahian, 2007; Mannion and Goddard, 2003).
- In the UK, annual reports on hospital performance include funnel plots, for example Figure 1.4.

1.3.6 A note of caution

Statistical inference is inextricably connected with uncertainty. Clinical registers naturally do not record all patients in all hospitals, due to time and cost constraints or on explicit request of the patient (e.g. privacy concerns). So in practice results are mostly based on a sample of n patients, where these results may still pertain to the whole patient population if this sample was randomly chosen, but not if for example (some) hospitals systematically avoid registering older patients with higher mortality risk.

Precision is decreased and bias may be induced when there are missing data (Rubin, 1976), e.g. when for some patients it is registered whether they smoke or not, while for others this is unknown. In Chapter 4 we discuss both complete case analysis and multiple imputation to handle missing data, comparing current practice and statistical preference (White and Carlin, 2010). To perform a complete case analysis, one assumes that the reason why values are missing is independent of the observed and unobserved data. In practice this assumption is rarely met and may then yield biased results for hospitals with a selective subset of complete records (Knol et al., 2010). Multiple imputation (Schafer, 1999) on the other hand assumes that given the observed data, the missingness mechanism is independent of the unobserved data, for example when age is more often missing for females whereby gender is completely observed. Missing values for age are then filled in, based on a parametric model that predicts age values in function of the observed data. So, whether imputed values for age are

unbiased, strongly relies on the correctness of the assumed parametric model. Special awareness is needed when the missing at random assumption is not met, for example when age is more often missing for older patients (Sterne et al., 2009). A sensitivity analysis may then explore how strongly results vary under the different missingness mechanisms (Robins et al., 2000).

In addition, benchmarking hospital performance is not a 'hard' science: Different techniques will give different results. Crossing a threshold should therefore not indicate high or low 'quality' but it may be useful to investigate reasons for the apparent discordance (Spiegelhalter, 2005b). Of course it would be pointless to collect and analyze all these data if no action follows. However, caution is needed when doing so. It has been shown, e.g. by Lilford et al. (2004) and Freeman (2002), that when results are used for funding or could even lead to closure, they often stimulate perverse reactions. One way is through data gaming, i.e. manipulating the data registration on purpose to upstage performance results, for example by keeping patients wait in ambulances to decrease the registered time between hospital admission and treatment start (Shaw et al., 2015). Another example is given in Figure 1.6 where some of the increasing trend in deaths coded as palliative will be due to an increasing accuracy of coding over the years, but it is highly likely that there is also an element of data gaming. Adverse reactions could even lead to worse clinical practice itself, e.g. when the number of surgical interventions for high-risk patients is reduced to avoid in-hospital deaths.

So, not only should there be paid close attention to developing valid statistical analysis but also to developing an organizational environment that encourages the constructive use of such information (Mannion and Goddard, 2003). Hospitals can use health outcomes data to evaluate how they are doing compared to their peers. This gives them a unique opportunity to learn from one another, to investigate and discuss its performance and next take whatever action seems necessary to improve the way they provide care. In this way, performance monitoring can indeed have a major beneficial effect on quality of care and patient outcomes (Gross et al., 2001): In Mehta et al. (2007) it is claimed that monitoring quality of care is likely to achieve larger reductions in death than any individual new therapy or drug.

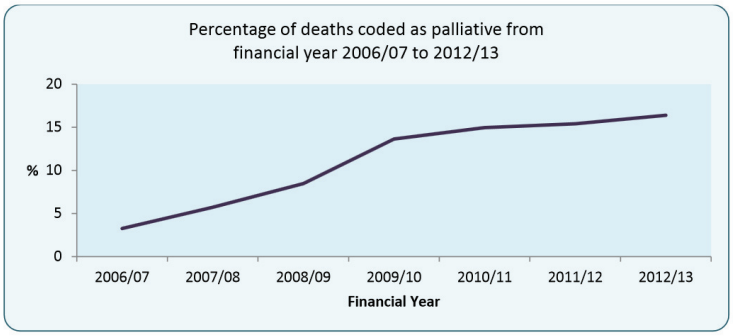


Figure 1.6: The percentage of deaths coded as palliative from financial year 2006/07 to 2012/13.

We can conclude with a citation from Lilford and Pronovost (2010): "The science still needs to mature, not only to improve the measurement of quality, but also to learn how to use the (inevitably imperfect) measurements so that they do more good than harm."

Example 1.5
In the UK, the Clinical Indicators Team (2015) remark that the reported standardized risks require careful interpretation and should be used in conjunction with other indicators and information from other sources (e.g. patient feedback, staff surveys and other similar material) that together form a holistic view of trust outcomes.

1.4 Technical Motivation and Outline

Having argued that hospital evaluations need careful statistical analysis which face many challenges, we give an overview of the statistical techniques in this thesis. We will focus on some deficiencies in the current methodology and evaluate the properties of candidate approaches in the following chapters.

Clinical registers often contain some hospitals with a small number of reg-

istered patients. It has been repeatedly shown that the popular mixed effects model may pull the performance of these hospitals towards the average, inflating the Type II error of not detecting outlying performance (Normand et al., 1997; Ash et al., 2012). Therefore we evaluate in **Chapter 2** to which amount Firth corrected estimates (Firth, 1993) may avoid such shrinkage and moreover reduce bias on the estimated directly standardized risk. A second concern we handle is the risk of model extrapolation when estimating e.g. the directly standardized risk, because it requires predicting for each patient how he/she would fare if the care level of a given hospital c applied, which is not observed. This is especially problematic when the patient mix differs substantially across hospitals, as illustrated at the top of Figure 1.7, where the smallest hospital suffers most from extrapolation. Standard regression methods may then hold biased results and underestimated uncertainty (Rubin, 1997). We investigate whether and how this could be remedied by weighting patients by the reciprocal of the probability to be treated in the observed hospital (Shahian and Normand, 2008). If this probability is extremely small, the end user will be warned for strong extrapolation through inflated variance estimates.

Common adjustments for differences in patient mix assume that the effect of the hospitals' care level on the outcome is constant over patient characteristics (Ohlssen et al., 2007b; Shahian and Normand, 2008). For example in model (1.3) the hospital effects ψ_1 and ψ_2 are assumed to be independent of the patient's age. This is violated when e.g. hospital 1 is specialized in care for the elderly, so that older patients receive much better care in hospital 1 compared to hospital 2, while for younger patients the care level is very similar in both hospitals. A visualization of the outcome regression models in such a situation is given in Figure 1.7 and will be discussed in more detail in **Chapter 3**. The topic of that chapter is to assess how common practice of ignoring such interactions may impact the bias and precision of directly and indirectly standardized risks, depending on the patient distribution across hospitals and the hospital size. This may help justify the common practice, especially in situations where it is simply prohibitive to allow for these interactions in the model because sufficient information is lacking in small hospitals, see for example Ash et al. (2012).

By adjusting for all potential confounders, outcome regression models aim to

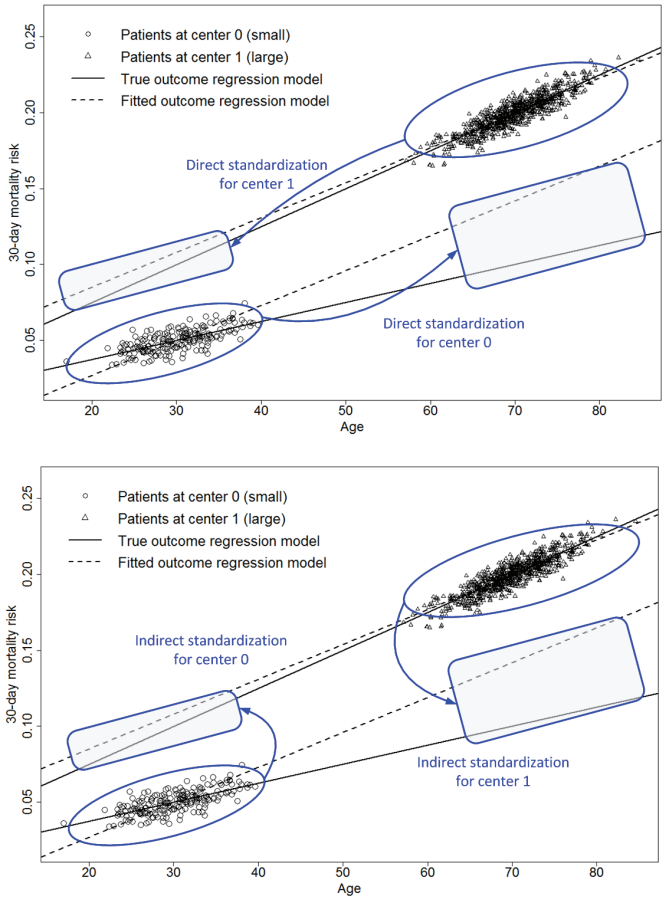


Figure 1.7: Extrapolation for the directly and indirectly standardized risk considering 2 centers (small or large center size). The 30-day mortality risk is estimated based on an outcome regression model without interaction between center and patient's age, while the true model includes a different center effect on 30-day mortality for varying ages.

come as close as possible to the true causal effect of hospitals on mortality (Greenland, 2008; Brookhart et al., 2010). However, registering all potential confounders is practically limited by the measurement time/cost and the perverse effect of more missing values (Shahian et al., 2007). Moreover, estimation strategies may no longer converge when too many covariates are included (Vansteelandt et al., 2010). In **Chapter 4** we investigate the use of stochastic search algorithms (De Beukelaer et al., 2015) to find the subset of confounders that minimizes the cross-validated error on the estimated individual or standardized risk. In doing so, we give the covariates a cost proportional to the measurement/registration effort or to the percentage of missing values, because missing values may imply serious loss of efficiency and even induce bias if the missingness assumption is not fulfilled (Sterne et al., 2009). We will also provide analytical guidelines that may help decide for which amount of missingness it is still beneficial to include a more informative covariate, rather than its surrogate which is completely observed.

We have also been working on a user-friendly R-package 'RiskStandard' that implements the estimation of the directly and indirectly standardized risks based on (Firth corrected) logistic regression models with fixed effects for the hospital and patient characteristics. The package allows for labeling hospitals based on these estimated risks, where the statistical and clinical significance levels are defined by the user. The implemented plot functions may help visualizing the quality of care in hospitals. This package is documented in **Chapter 5** and can be downloaded from www.cvstat.ugent.be.

CHAPTER 2

On Shrinkage and Model Extrapolation in the Evaluation of Clinical Center Performance

This chapter is based on the following paper: Varewyck, M., Goetghebeur, E., Eriksson, M., and Vansteelandt, S. (2014). “On shrinkage and model extrapolation in the evaluation of clinical center performance," *Biostatistics*, 15(4): 651-664.

Summary

We consider statistical methods for benchmarking clinical centers based on a dichotomous outcome indicator. Borrowing ideas from the causal inference literature, we aim to reveal how the entire study population would have fared under the current care level of each center. To this end, we evaluate direct standardization based on fixed versus random center effects outcome models that incorporate patient-specific baseline covariates to adjust for differential case-mix. We explore fixed effects regression with Firth correction and normal mixed effects regression to maintain convergence in the presence of very small centers. We moreover study doubly robust fixed effects regression to avoid outcome model extrapolation. Simulation studies show that shrinkage following

standard mixed effects modeling can result in substantial power loss relative to the considered alternatives, especially for small centers. Results are consistent with findings in the analysis of 30-day mortality risk following acute stroke across 90 centers in the Swedish Stroke Register.

2.1 Introduction

In recent years, the interest in profiling hospital performance has grown among different stakeholders including government and health insurers, hospitals and clinicians, and last but not least the patients. Health care quality thus deserves careful statistical analysis yielding relevant and relatively simple measures with clear interpretation for hospital evaluation.

In this article, we focus on statistical methods to estimate center performance on a binary quality indicator such as 30-day mortality. Causal inference methods will be adopted to adjust for measured confounding by differential patient mix (e.g. initial disease status, age). This is important (DeLong et al., 1997; Austin et al., 2003) as centers treating more severely ill patients tend to have higher mortality irrespective of treatment quality. Most literature uses indirect standardization to adjust for patient mix (Spiegelhalter, 2005a; Shahian and Normand, 2008). This involves contrasting the observed average quality outcome in each center with what it would have been for their patients if 'the average level of care over all centers' applied. This is particularly helpful for policy makers when deciding where to best spend resources for quality improvement. However, when centers are expected to provide good health care on the overall patient population, directly standardized outcomes may be of greater interest. This potential full population risk in each center will be our focus. It makes us consider how the entire study population would have fared under the current level of care of each center.

Random effects models which incorporate patient-specific baseline covariates are routinely applied for indirect standardization (Ohlssen et al., 2007a), and can also be used for direct standardization. The main advantage of these models is that they severely reduce the effective model dimension, thereby avoiding

problems of overfitting. However, two main shortcomings deserve more in depth study: shrinkage and model extrapolation. First, estimates for small centers may shrink severely toward the population mean, resulting in bias and power loss for these centers (Normand et al., 1997). This is a major concern because the quality of care in small centers is sometimes questioned, in view of their potentially more limited surgical experience or medical infrastructure (Saposnik et al., 2007). Fixed effects models are no viable substitute in settings encountering many centers with often small numbers of registered patients, where this method suffers from bias and convergence problems (Neyman and Scott, 1948). In this article, we will investigate whether this limitation can be overcome via the Firth correction for fixed center effects models (Firth, 1993). Second, when case-mix differs severely between centers, results from the default fixed and random effects models can become very sensitive to model misspecification which is hard to detect (Rubin, 1997). We aim to overcome this using doubly robust methods (Robins et al., 2007) that build on a fixed center effects model (with Firth correction) but utilize inverse weighting by the so-called propensity score (Shahian and Normand, 2008), which is the probability of being treated in the observed center based on patient characteristics.

These statistical methods will be compared in terms of their support for correct detection of low and more importantly high risk centers. For this purpose, we will adapt the decision criterion suggested by Normand et al. (1997) to the framework of direct standardization. Specifically, we will seek solid statistical evidence of a clinically relevant difference between the potential full population risk from a given center and the observed population risk.

Comparisons are made in two case studies: a simulation study on quality insurance for rectal cancer treatment in Belgium and an analysis of quality of care data from the Swedish Stroke Register (Asplund et al., 2011). They reflect markedly different settings with chronic versus acute illness, with major versus more limited differences in case-mix, with small versus larger center sizes, and with a limited versus rich set of patient covariates.

2.2 Profiling Center Performance: Framework

Throughout the paper, C is a random variable indicating in which center the patient was actually treated ($C = 1, \ldots, m$) and $\mathbf{L}$ denotes the vector of patient-specific baseline characteristics such as gender and initial disease status. The methods below focus on 30-day mortality Y, but can easily be extended to a continuous or categorical outcome.

2.2.1 Direct versus indirect standardization

Direct standardization aims to infer the potential full population risk for each center c: the risk that would be realized if all patients under study were to experience the care level of that given center c, irrespective of where they were actually treated. We denote this by $E\{Y(c)\}$, where $Y(c)$ indicates the potential outcome for given patient if treated at the care level of center c. A main feature of direct standardization is that the patient mix used for comparison is a common set of subjects. As such, center comparisons are based on their current performance in the extended patient population, where the extent of extrapolation from each center's own patient population depends on how case-mix differs between centers. This approach may thus evaluate a center's performance based on patients it is not likely to treat.

In contrast, indirect standardization focuses on what a center achieves for its own patient mix. For instance, the frequently used standardized mortality ratio (SMR) takes the ratio of the center's observed risk and the expected risk if these patients would experience the average care level across all centers where center performance levels were uniformly distributed, i.e.

$$\text{SMR} = \frac{E\{E(Y|\mathbf{L}, C = c)|C = c\}}{m^{-1}\sum_{c^*=1}^{m} E\{E(Y|\mathbf{L}, C = c^*)|C = c\}} = \frac{E\{Y(c)|C = c\}}{m^{-1}\sum_{c^*=1}^{m} E\{Y(c^*)|C = c\}}, \quad (2.1)$$

(DeLong et al., 1997; Shahian and Normand, 2008). When a difference is taken instead of a ratio, the name 'excess risk' is used (Goetghebeur et al., 2011). Indirect standardization thus aims to answer the question: 'How would the risk in a given center change if its patients were to experience the average risk across all centers?'.

	Mortality risk (no. patients)		Indirect stand.		Direct stand.
	L = low	L = high	SMR	Excess Risk	
center 1	1% (900)	10% (100)	0.8382	−0.0037	0.0670
center 2	1% (100)	10% (900)	0.9349	−0.0063	0.0670
center 3	2% (100)	12% (900)	1.1301	0.0127	0.0833

Table 2.1: Artificial example comparing center performance based on indirect and direct standardization. For the three centers patient-specific mortality risks and patient mix (no. patients) are given per level of the covariate L, indicating low or high baseline severity.

Such contrast with the average risk across all centers can be limiting when this reference deviates from what is ideally targeted.

Direct and indirect standardization extrapolate observations to a general population or a general care level respectively. This may result in different comparisons, as illustrated in Table 2.1 where centers 1 and 2 have the same patient-specific mortality risks, but differ in patient mix. Following indirect standardization, these centers are classified as having different performance because their patient mix differs. Results moreover depend on whether indirect standardization is based on SMRs or excess risks, because small absolute differences can result in large relative differences and vice versa. Therefore, when indirectly standardized outcomes are of interest we would recommend excess risks emphasizing the possibly large extrapolation of center performance. The directly standardized risks on the other hand detect equal quality of care in centers 1 and 2, because of their equal patient-specific mortality risks. They also allow for direct comparison with the overall risk of 7.02%, but may involve serious extrapolation when the patient population of that center differs substantially from the overall population.

2.2.2 Decision criterion for labelling centers

In Section 2.3, we will compare statistical methods for direct standardization in terms of correctly detecting low and more importantly high risk centers. Therefore, following a proposal first introduced in a Bayesian context (Normand et al., 1997), we will classify a center as low/high risk if the data provide sufficient

evidence that the potential risk $E\{Y(c)\}$ exceeds a benchmark relative to the population average risk $E(Y)$. For this purpose, we will develop estimators $\hat{E}\{Y(c)\}$ for the potential risk $E\{Y(c)\}$ (see Section 2.3) and then classify a center as low risk if

$$\hat{E}\{Y(c)\} + z_k \times \mathrm{sd}\left(\hat{E}\{Y(c)\}\right) < (1-\lambda)\, E(Y) \tag{2.2}$$

or as high risk if

$$(1+\lambda)\, E(Y) < \hat{E}\{Y(c)\} - z_k \times \mathrm{sd}\left(\hat{E}\{Y(c)\}\right). \tag{2.3}$$

Here, λ expresses a clinically meaningful tolerance level (e.g. 20%) indicating how much the center-specific potential risk is allowed to depart from the current population average risk $E(Y)$. The latter can be estimated by the sample average of observed risks or be replaced by a reference standard if objective benchmarks (e.g. national guidelines) are available. In practice such envisaged reference is likely to steer the choice of λ once $E(Y)$ is known or has been estimated. Further, z_k is the $k \times 100$th percentile of the standard normal distribution, so k (e.g. 0.75) expresses the degree of statistical evidence required before flagging a center as low/high risk.

The previous criterion has close links to the often used frequentist criterion (DeLong et al., 1997) whereby a center c is classified as low/high risk if the estimated 95% confidence interval for its potential full population risk excludes the population average risk $E(Y)$:

$$E(Y) \notin \left[\hat{E}\{Y(c)\} \pm z_{0.975} \times \mathrm{sd}(\hat{E}\{Y(c)\})\right]. \tag{2.4}$$

See Shahian and Normand (2008) for a related Bayesian criterion. This corresponds with (2.2) and (2.3) if $\lambda = 0$ and $k = 0.975$. A key drawback of this criterion is that it disregards clinical significance. In particular, large centers are virtually guaranteed to exclude the population average risk and thus will nearly always be labelled statistically significant low/high risk centers.

Pure ranking based on the estimated potential risk is dangerous as it is oblivious to a clinical appreciation of differences between centers as well as to uncertainty. Large differences in ranks may correspond with small clinical differences

and vice versa. Moreover, uncertainty around the ranking is often substantial (especially for small centers) and confidence intervals may tend to overlap for different centers (Spiegelhalter, 2005a). Although ranking based on the estimated probability of exceeding performance can be considered (Normand et al., 1997), we believe its inherent property of masking the size of differences in center performance demands careful interpretation involving additional information; it will therefore not be considered in this paper.

2.3 Regression Methods

We will now discuss different methods to estimate the potential full population risk $E\{Y(c)\}$ in each center $c = 1, \ldots, m$. Let n be the total sample size. Throughout we assume that the patient-specific covariates $\mathbf{L}$ are sufficient to adjust for confounding of the center-outcome effect, so that $Y(c) \perp\!\!\!\perp C|\mathbf{L}$ for all c (Hernán and Robins, 2006b). Under this assumption, we have that

$$E\{Y(c)\} = E\{E(Y|\mathbf{L}, C = c)\}. \tag{2.5}$$

2.3.1 Normal mixed effects model

We will first focus on outcome regression models which postulate that in each center c:

$$E(Y|\mathbf{L}, C = c) = \text{expit}\left(\mathbf{L}'\boldsymbol{\beta} + \psi_c\right) = f(\mathbf{L}, c; \boldsymbol{\beta}, \boldsymbol{\psi}), \tag{2.6}$$

where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_m)$ are the center effects. For convenience, we here constrain the effects $\boldsymbol{\beta}$ of patient-specific covariates on outcome to be equal across all centers, but this can in principle be checked (size permitting) or relaxed by including interactions with center. Once estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}})$ for $(\boldsymbol{\beta}, \boldsymbol{\psi})$ have been obtained, it follows from (2.5) that the potential full population risk can be estimated as (Hernán and Robins, 2006b):

$$\hat{E}\{Y(c)\} = \frac{1}{n}\sum_{i=1}^{n} \text{expit}\left(\mathbf{L}_i'\hat{\boldsymbol{\beta}} + \hat{\psi}_c\right) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{L}_i, c; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}). \tag{2.7}$$

The evaluation of center performance is often based on Bayesian normal

mixed effects models (ME). These augment model (2.6) with a normal random effects distribution

$$\psi_c \sim N(\mu_\psi, \sigma^2_\psi), \quad c = 1, \ldots, m, \tag{2.8}$$

which assumes centers to be exchangeable in the sense that any a priori information on the relative ordering or grouping of center effects is ignored (Ohlssen et al., 2007a). Here, μ_ψ is the common mean and σ_ψ the standard deviation of the center effects ψ_c, which we assume to have independent hyperpriors. Moreover, assuming that the center effects are a priori independent of the effects of patient characteristics, the joint posterior distribution of this two-level Bayesian approach is of the form:

$$p(\beta, \psi, \mu_\psi, \sigma^2_\psi | \mathbf{y}, \mathbf{L}, \mathbf{C}) \propto \prod_{i=1}^{n} p(y_i | \beta, \psi, \mathbf{L}, \mathbf{C})\, p(\psi | \mu_\psi, \sigma^2_\psi)\, p(\beta)\, p(\mu_\psi)\, p(\sigma^2_\psi). \tag{2.9}$$

This posterior is estimated using a Markov chain Monte Carlo (MCMC) algorithm, which provides values for the model parameters (β, ψ) in each step. These are subsequently used to evaluate (2.7), thereby enabling us to estimate the posterior distribution of $E\{Y(c)\}$.

An advantage of the Bayesian over a frequentist approach based on empirical BLUPs (Robinson, 1991) is that by using MCMC algorithms one can directly obtain posterior estimates and variances of even complicated transformations such as (2.7) without the need for large sample justifications, provided a sufficient number of MCMC iterations are run (O'Brien and Dunson, 2004). Prior information can be incorporated in Bayesian models through an informative prior distribution. When no such information is available, a normal distribution with large variance as non-informative prior on center level may still be hard to justify as the center effects could follow a longer-tailed distribution such as the Student's t or even an asymmetric distribution. In particular, choosing a normal prior may shrink estimated center effects towards the center population mean μ_ψ, especially for very small centers (Normand et al., 1997). Severe shrinkage may be problematic when reporting individual feedback to the centers, because identification of centers with deviating performance is especially important. The amount of shrinkage is related to the choice of prior, but judging the plausibility

of a normal prior is difficult because it refers to center effects on the logit scale.

2.3.2 Reducing shrinkage

Clustered normal mixed effects model.

The mixture model of Ohlssen et al. (2007a) forms a first approach considered to reduce shrinkage. This involves assigning the m centers to a chosen number $K < m$ of clusters with each their own normal random effects distribution. The model for the center level effects thus becomes

$$\psi_c \sim N(\mu_k, \sigma_k^2) \quad \text{with unknown probability } p_k, \;\; k = 1, \ldots, K.$$

It thus assigns each center c to cluster k with probability p_k and subsequently draws a random center effect from the normal distribution of the 'cluster k population' with cluster mean μ_k and variance σ_k^2 within the cluster. In this process, we let each center have an a priori equal probability of belonging to each cluster without the size or performance of the clusters being predefined. When a priori knowledge of clustered center performance is available (e.g. when large centers are expected to have better facilities resulting in better performance (Saposnik et al., 2007)), it can be incorporated by giving centers a larger prior probability for specific clusters.

Fixed effects logistic regression model with Firth correction.

Shrinkage can alternatively be reduced using maximum likelihood estimation for the fixed effects (FE) logistic regression model (2.6). However, because of overfitting the resulting estimator (2.7) may behave erratically when there are centers with few events: besides convergence problems, there may be substantial finite sample bias and large variance (Peduzzi et al., 1996). The ME approach of Section 2.3.1 accommodated this by imposing a normal distribution on the center effects. Here, we will consider the Firth corrected FE method instead.

Firth correction (Firth, 1993) reduces the $O(n^{-1})$ bias of ordinary maximum likelihood estimators to the order $O(n^{-2})$ by maximizing the penalized likelihood

function

$$L^*(\boldsymbol{\beta},\boldsymbol{\psi}) = L(\boldsymbol{\beta},\boldsymbol{\psi})|I(\boldsymbol{\beta},\boldsymbol{\psi})|^{1/2}, \tag{2.10}$$

instead. Here, $|I(\cdot)|$ denotes the determinant of the Fisher information matrix of $(\boldsymbol{\beta},\boldsymbol{\psi})$ and $L(\cdot)$ is the ordinary likelihood function. Since $|I(\boldsymbol{\beta},\boldsymbol{\psi})|^{1/2}$ equals Jeffreys' invariant prior, Firth correction is equivalent with penalization of the likelihood by Jeffreys' prior (Kosmidis and Firth, 2009), suggesting that Firth corrected maximum likelihood estimates are also subject to shrinkage. However, Jeffreys' prior is invariant under reparameterization and has the key feature of being non-informative. The latter, coupled with its defining bias reduction property, implies that it may result in less shrinkage compared to the use of a normal prior. There is evidence that it may also perform better in terms of other properties such as finiteness of the estimator and coverage of confidence intervals (Kosmidis and Firth, 2009). We will investigate this in our setting through simulations in Section 2.4. In the Appendix (Section 2.A.1 and 2.A.2), we give additional detail on the Firth correction and show how to estimate the asymptotic variance of the resulting estimate of $E\{Y(c)\}$.

2.3.3 Accounting for model extrapolation: Doubly robust propensity scores method

All previous methods suffer from a risk of extrapolation as they require predicting for each patient how he/she would fare if the care level of a given center c applied. When case-mix differs across centers, especially when there are strong confounders, such extrapolation may not be justified. Even models that seem to fit the observed data well may then be misspecified and imply serious model extrapolation, resulting in bias and underestimated uncertainty (Rubin, 1997). This is illustrated in the Appendix (Figure 2.4) where we consider two centers with strongly differential case-mix: one center has patients older than 60 and the other does not. We find strong model extrapolation that may not get reflected in standard errors, so the user is left without warning (Rubin, 1997). Similar concerns are warranted for standard indirect standardization methods as these extrapolate stratum-specific center effects to the patients of each given center.

Inverse probability weighting via propensity scores (PS) avoids extrapolation by not relying on outcome models. For a given patient i the PS are defined as the vector of probabilities to belong to each given center c on the basis of his/her baseline characteristics $\mathbf{L}_i$. In practice, such PS can be estimated by fitting a multinomial regression model:

$$P(C_i = c|\mathbf{L}_i) = g(\mathbf{L}_i, c; \boldsymbol{\gamma}, \boldsymbol{\delta}) = \begin{cases} \frac{1}{1+\sum_{j=2}^{m} \exp(\mathbf{L}_i'\boldsymbol{\delta}_j + \gamma_j)} & c = 1 \\ \frac{\exp(\mathbf{L}_i'\boldsymbol{\delta}_c + \gamma_c)}{1+\sum_{j=2}^{m} \exp(\mathbf{L}_i'\boldsymbol{\delta}_j + \gamma_j)} & c \neq 1, \end{cases} \tag{2.11}$$

where C_i indicates the center where patient i was treated. Parameter estimators $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})$ can be obtained via maximum likelihood, so by solving the following set of estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} 1 \\ \mathbf{L}_i \end{pmatrix} \left\{ I(C_i = c) - g(\mathbf{L}_i, c; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}) \right\} = \mathbf{0} \qquad c = 2, \ldots, m. \tag{2.12}$$

We can now estimate $E\{Y(c)\}$ as in (2.7), but using a weighted regression to fit the fixed effects model (2.6), with weights equal to one over the PS of the observed center $g(\mathbf{L}_i, C_i; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})$. That this works can be seen because the weighted regression of the FE model sets

$$\frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \mathbf{L}_i \\ I(C_i = 1) \\ \vdots \\ I(C_i = m) \end{pmatrix} \frac{1}{g(\mathbf{L}_i, C_i; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})} \left\{ Y_i - f(\mathbf{L}_i, C_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}) \right\} = \mathbf{0}. \tag{2.13}$$

This enables us to rewrite

$$\begin{aligned} \hat{E}\{Y(c)\} &= \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{L}_i, c; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[f(\mathbf{L}_i, c; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}) + \frac{I(C_i = c)}{g(\mathbf{L}_i, c; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})} \left\{ Y_i - f(\mathbf{L}_i, c; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}) \right\} \right] \quad (2.14) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[\frac{I(C_i = c) Y_i}{g(\mathbf{L}_i, c; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})} + \left\{ 1 - \frac{I(C_i = c)}{g(\mathbf{L}_i, c; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})} \right\} f(\mathbf{L}_i, c; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}) \right]. \quad (2.15) \end{aligned}$$

These expressions show how the resulting estimator of $E\{Y(c)\}$ is doubly robust, i.e. unbiased (in large samples) if either the outcome or the PS model holds, but not necessarily both (Robins et al., 2007). Indeed, the second term in (2.14) and (2.15) has population mean zero if respectively the outcome or the PS model is correctly specified. Furthermore, the remaining term has mean $E\{Y(c)\}$ under those respective assumptions. This double robustness property is attractive because it allows for misspecification of the FE model if the PS is modelled correctly. It thus in particular offers partial protection against false omission of interactions between center and patient characteristics. When patient mix is drastically different between centers, the resulting lack of information gets exhibited in large standard errors.

Small centers can give problematically small estimated PS values, especially when its patient case-mix is very different from that of other large centers. We therefore stabilize the PS for each center by dividing all estimates $g(\mathbf{L}_i, c; \hat{\gamma}, \hat{\boldsymbol{\delta}})$ by the proportion of patients at that center. This stabilization does not affect the consistency nor the double robustness property of $\hat{E}\{Y(c)\}$. These properties are also not affected by applying the Firth correction when fitting the outcome model (2.6) by weighted regression, because this is a finite-sample correction. In the Appendix (Section 2.A.3) we give details on how the asymptotic variance of $\hat{E}\{Y(c)\}$ is estimated.

2.4 Results

2.4.1 Simulation study application: Belgian Colorectal Cancer register

The Belgian cancer register collected data on colorectal cancer diagnosis and follow-up between 2006 and 2010. This fairly young register with voluntary participation has a national coverage of about 30%. We will consider a total of 2355 patients treated in 63 centers and examine a binary outcome quality indicator with 22% events on average. Causal inference methods on this dataset were introduced by Goetghebeur et al. (2011), with descriptive statistics showing

substantial heterogeneity in case-mix.

We will compare the performance of the normal ME, clustered normal ME, Firth corrected FE and doubly robust PS methods (Section 2.3) via simulation experiments that reflect the structure of these data. Comparisons are based on the power and type I error of center classification following the profiling technique described in section 2.2.2 with $\lambda = 20\%$ and $k = 0.75$. For example, the Power to detect High is the probability of classifying a center as having high risk, given that its 'true' classification is high. The balance between type I error and power is determined by the values of the clinical (λ) and statistical (k) tolerance levels and the distribution of true alternatives, which are fixed here. Details on the simulation study are given in the Appendix (Section 2.B.1). Results are shown in Table 2.2 and Figure 2.1, where it can be seen that the power to detect low risk centers is generally smaller than the power to detect high risk centers. This is expected because of the lower number of events and closeness to the boundary for low mortality risk. Because of shrinkage, the normal ME method has very low power and appears unable to detect many of the low/high risk centers. Mistakenly classifying centers as low/high mortality risk is equally rare for these methods.

Interestingly, the clustered normal ME method is not performing better. This blind clustering, i.e. irrespective of center characteristics, thus requires considerable computing effort with no pay back at the considered sample size. While similar in terms of power, the doubly robust PS method makes more Type I errors than the FE method because of the lower coverage of the nevertheless on average wider confidence intervals. Although the doubly robust PS method has potential value in realistic settings with strong confounding because of its robustness against model misspecification, confidence intervals with better finite sample performance are needed before routine application can be recommended. The percentage of centers that are correctly classified is similar for all methods. For the ME methods, unlike for the other methods, this is the result of classifying nearly all centers as 'accepted' (see Figure 2.1).

In the Appendix (Section 2.B.1) we provide additional simulation results: Applying the Firth correction did not severely influence the results of the uncorrected FE methods, but ensured convergence in the presence of very small

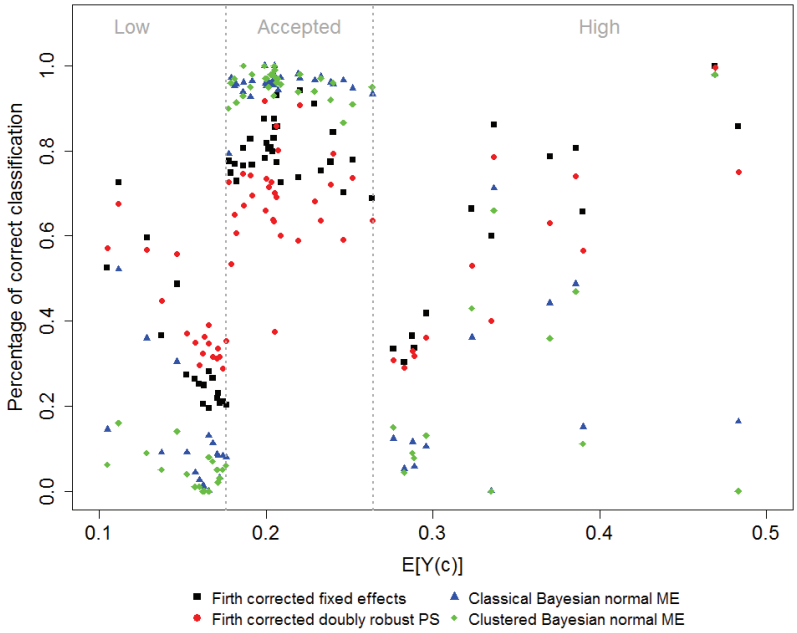


Figure 2.1: Percentage of correct classification against true potential full population risk for each center and regression method, based on the simulation study mimicking the setting of the Belgian colorectal cancer register. The vertical gray lines indicate the clinical decision limits $(1-\lambda)E(Y)$ *and* $(1+\lambda)E(Y)$*, with* $\lambda = 0.20$.

	Classical normal ME	Clustered normal ME	Firth corrected FE	Firth corrected doubly robust PS
Power (%)				
to detect High	28.8	31.2	61.5	53.9
to detect Low	12.3	5.4	32.0	39.9
Type I error (%)				
Low as High	0.1	0.3	2.4	6.3
Accepted as High	1.2	2.0	8.6	11.8
Type I error (%)				
High as Low	0.1	0.1	0.9	2.9
Accepted as Low	3.2	2.1	11.2	19.1
Coverage of 95% *CI*				
for $E\{Y(c)\}$ *(%)*	95.4	87.5	93.9	89.0
Classification				
% High (22%)	7.2	8.0	16.5	17.9
% Low (30%)	6.0	3.0	16.0	22.2
% correct (L-A-H)	59.6	57.7	62.4	58.1

Table 2.2: Center classification (Low/High risk or Accepted) based on 1000 *simulations for each regression method. The true percentage of low and high risk centers is respectively* 30% *and* 22%*.*

centers. Doubling the sample size especially increased the accuracy of the ME methods. When only one (outcome or PS) model is misspecified we found some evidence favouring the doubly robust PS method over the FE method.

2.4.2 Analysis of the Swedish Stroke Register

Riksstroke (http://www.Riksstroke.org) is a national quality register for acute stroke, collecting data from all 90 Swedish hospitals. It has an estimated coverage of the total stroke population between 80% and 90%, but there is considerable variation in coverage between hospitals. The setting is conceptually different from the cancer register, since acute stroke is treated urgently, mostly at the nearest center. The register is linked with a socio-economic database at Statistics Sweden and contains 149 778 patients with first stroke between 2001 and 2009. Centers are compared on 30-day mortality, applying the classical normal ME, the FE and the doubly robust PS method, without Firth correction as centers are no longer problematically small.

Because of convergence issues with multinomial models, we build a separate

logistic regression model $P(C = c|\mathbf{L}) = \text{expit}(\mathbf{L}'\boldsymbol{\delta}_c + \gamma_c)$ per center c and estimate the PS for individual i treated at center c as:

$$\frac{P(C_i = c|\mathbf{L}_i)}{\sum_{j=1}^{m} P(C_i = j|\mathbf{L}_i)} = \frac{\text{expit}(\mathbf{L}_i'\boldsymbol{\delta}_c + \gamma_c)}{\sum_{j=1}^{m} \text{expit}(\mathbf{L}_i'\boldsymbol{\delta}_j + \gamma_j)}. \tag{2.16}$$

The classical ME method is applied separately to the cluster of small (< 1000 patients), medium (1000 to 2000) and large (> 2000) centers, to reduce the effects of shrinkage and to avoid convergence problems due to the large data size. The potential full population risk for this method is then based on the cluster-specific parameter estimates applied to the total population.

While in general few data are missing, records on education and smoking are missing for respectively 20.8% and 12.8% of the participants. Certain patient characteristics, like smoking status, are more likely missing for patients who were unconscious upon admission and education level is more often unknown for elderly patients. We fit complete-case regression models as they allow for missingness in the covariates to depend on the covariates themselves, so long as there is no residual dependence on the outcome (White and Carlin, 2010); it moreover avoids the need for modelling the distribution of those covariates. They will therefore return center effects that are unbiased under fairly minimal assumptions. Since the proposed estimators standardise these effects to the same reference population, selective missingness is not expected to distort comparison of $E\{Y(c)\}$ between centers, although it may yield underestimates of $E\{Y(c)\}$ for each center, as suggested by the comparison of complete cases (n = 101 051) versus all cases in the Appendix (Section 2.B.2). The Appendix further specifies the covariates that were included in the PS and outcome regression models. It shows that in general case-mix does not differ much across centers, except for considerable variation with respect to treatment for high blood pressure, education level and time of admission.

The fewer centers classified as low/high risk by the normal ME method, are also found by the other methods. This can be seen in Figure 2.2 which displays the observed center-specific risk $\hat{E}(Y|C = c)$ (which does not depend on the analysis method) versus the model based potential full population risk $\hat{E}\{Y(c)\}$. The

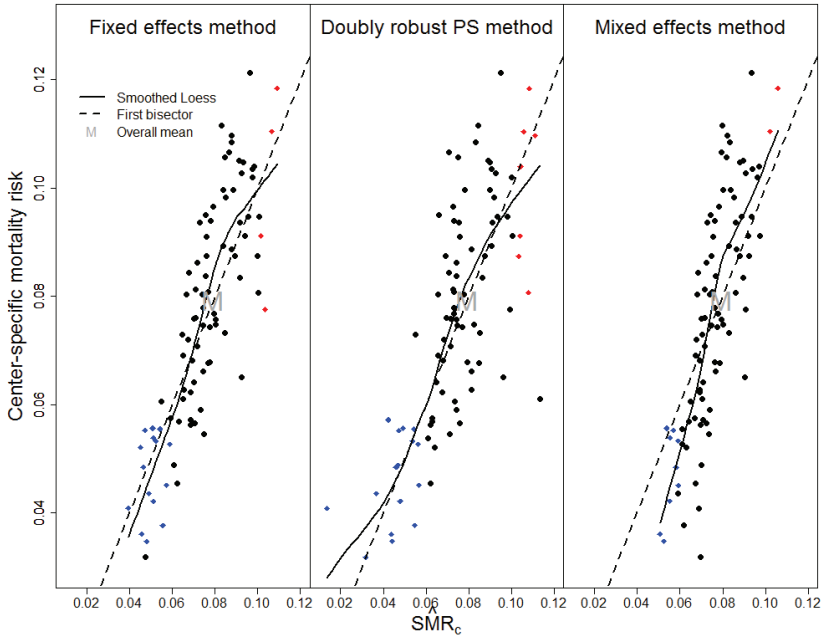


Figure 2.2: The observed center-specific risk versus the potential full population risk for all 90 centers of Riksstroke for the FE method, the doubly robust PS method and the classical normal ME method, distinguishing between low (filled circle) and high (filled triangle) mortality risk.

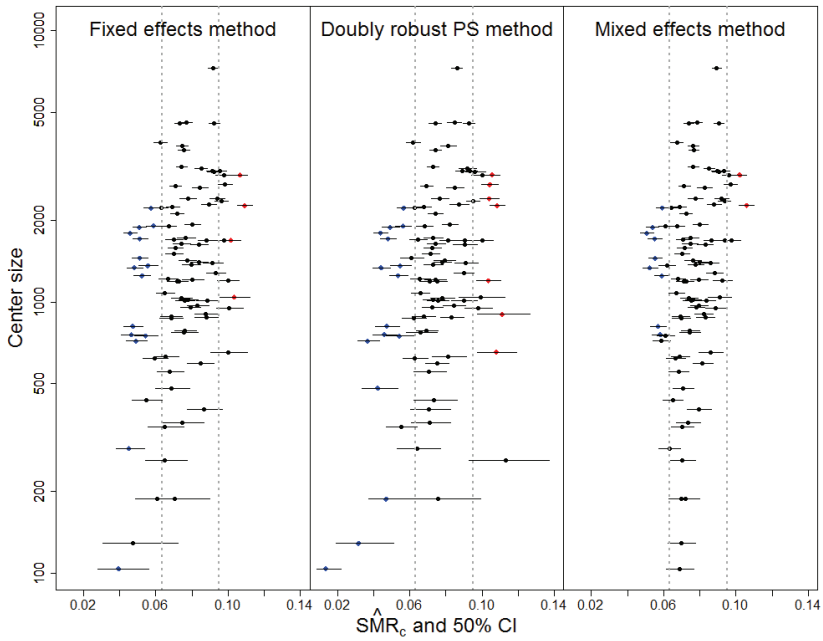


Figure 2.3: The potential full population risk and corresponding 50% confidence interval for all 90 centers of Riksstroke for the FE method, the doubly robust PS method and the classical normal ME method. The vertical gray lines indicate the clinical decision limits $(1-\lambda)\hat{E}(Y)$ *and* $(1+\lambda)\hat{E}(Y)$*, with* $\lambda = 0.20$*, which are used for classification of low and high (filled circles) mortality risk centers.*

ME results show the smallest range of potential full population risk with most attenuation towards the overall risk, i.e. more shrinkage. Both FE methods classify mostly the same centers as low risk, although the doubly robust PS method classifies two additional centers as high risk (Figure 2.2). Because these two centers may be only borderline statistically significant, we examine the uncertainty in terms of a 50% CI on the potential full population risk (Figure 2.3). We found good discrimination of low and high mortality risk in general, but the 50% CI of some centers is close to the clinical decision limit where it becomes difficult to judge. In that case, balanced testing could help to find an optimal combination of the null and the alternative (Moerkerke and Goetghebeur, 2006). Figure 2.3 shows better accuracy of $\hat{E}\{Y(c)\}$ with increasing center size, except for centers with close to zero events. Surprisingly, high mortality risks are observed mostly for medium to large centers while low mortality risks are especially detected for small to medium centers. This may partly be due to selectivity as it is known that patients dying early are less likely to be recorded in this Riksstroke register which could potentially happen more frequently in smaller centers. Since size is based on the number of registered patients, a center with low coverage may come out as smaller with better performance. This underlines again the importance of complete coverage.

For the doubly robust method we observe wider confidence intervals especially for the small centers. This may be related to the generally lower efficiency of this method, but also be more honestly reflecting the uncertainty on the potential risk estimates. Unlike the other two methods, the ME method did not classify any of the very small centers as low risk. This is due to shrinkage to which especially the smallest centers are very sensitive.

2.5 Discussion

We have proposed and compared approaches to evaluate the performance of clinical centers via direct standardization. This involves comparing centers in terms of the potential risk if the full study population were treated at the current level of care of the given center. A key feature is that the evaluation of all centers

is based on the same reference population, while each center will have treated a subset. Especially when centers can be chosen freely, this cuts bias out of the current center performance. Alternatively, indirect standardization, which is more widely used (Shahian and Normand, 2008; Austin et al., 2003), evaluates each center on its own patient population and is of particular relevance when centers tend to differ in their patient mix. Both standardizations have their virtues and in future work we will develop similar analysis strategies for indirect standardization.

We have compared three statistical regression methods for direct standardization based on random or fixed center effects, the latter in combination with the Firth correction or weighting by the reciprocal of the PS to be treated in the observed center. Our primary focus on frequentist methods was motivated by the fact that they are less computer-intensive and avoid the need to specify prior distributions about which no information was available in our case studies. A crucial and unverifiable assumption for all considered methods is that the included set of baseline patient characteristics is sufficient to adjust for confounding of the center-outcome effect. This drives the variable selection at the design stage of disease registers. In case of violations, results will be biased and one may want to consider other methods such as those using instrumental variables (Hernán and Robins, 2006a).

In the first case study, we found that shrinkage following traditional ME modeling results in substantial power loss compared to the suggested alternatives, especially for small centers. Although we used direct standardization, these findings correspond to those observed for indirect standardization in Austin et al. (2003). The Firth corrected FE model as well as the doubly robust PS method recovered power, while maintaining convergence in the presence of very small centers. In the second case study, shrinkage was still present under the normal ME model, but disappeared using the Firth correction (see Appendix, Section 2.B.2). As a result, fewer centers were classified as low/high risk under the random compared to the fixed center effects models.

In the simulation study the Firth corrected FE method outperformed the doubly robust PS method, although differences were relatively minor. While routine application of the doubly robust PS method in its current form is not

recommended, it may be of potential interest in settings with strongly differential case-mix (Shahian and Normand, 2008). In such settings, standard variable selection procedures for outcome regression models which force the center effects into the model have a tendency, as a result of multicollinearity, to exclude patient characteristics that are strongly correlated with center choice so that their effects get attributed to differences between centers. This may in turn yield model extrapolation with biased and misleadingly precise center effect estimates (Vansteelandt et al., 2010). The doubly robust PS method helps to protect against this. By also modeling the effect of patient characteristics on the center choice, stronger predictors of center choice are potentially more likely to be picked up in variable selection procedures (Hocking, 1976). In future work, it will therefore be of interest to evaluate how the considered methods perform when combined with variable selection. Double robustness moreover protects against misspecification of the outcome model when the model for center choice is correct. It thereby lessens the concern for violation of the assumption of equal covariate effects across centers.

2.A Technical Appendices

Software in the form of R-code is available at
https://github.com/mmvarewy/Biostatistics_2014.

2.A.1 Firth correction

Maximum likelihood estimation for the parameter $\boldsymbol{\theta} := (\boldsymbol{\beta}, \psi)$ indexing a logistic regression model amounts to solving the equations

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = u(\boldsymbol{\theta}) = \mathbf{0}, \tag{2.17}$$

where $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ is the log likelihood function and $u(\boldsymbol{\theta})$ is the score function for a single observation. Firth (1993) defines a modified score function

$$u^*(\boldsymbol{\theta}) = u(\boldsymbol{\theta}) - I(\boldsymbol{\theta}) b(\boldsymbol{\theta}), \tag{2.18}$$

where $I(\boldsymbol{\theta}) = -\frac{\partial u(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is the Fisher information matrix for a single observation and $b(\boldsymbol{\theta})$ is the $O(n^{-1})$ bias of the resulting maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. Therefore, the solution θ_F to $u^*(\boldsymbol{\theta}) = \mathbf{0}$ removes the $O(n^{-1})$ bias of $\hat{\boldsymbol{\theta}}$.

A simple interpretation of equation (2.18) is that it specifies an amount of bias to be introduced into the score function in order to remove the leading term in the asymptotic bias of $\hat{\boldsymbol{\theta}}$. The introduced bias function $-I(\boldsymbol{\theta}) b(\boldsymbol{\theta})$ is $O(1)$ as $n \to \infty$, an order of magnitude smaller in probability than $u(\boldsymbol{\theta})$ itself (Firth, 1992). The Firth approach eliminates the $O(n^{-1})$ bias of the maximum likelihood estimate by adjusting the score function, rather than the estimate itself. An advantage is that the definition of the Firth corrected estimates does not depend on the existence of the non-corrected estimates, of which some components may be infinite in binomial problems (Firth, 1992).

Simulation studies in different settings for logistic regression in Heinze (2006) show promising results for the Firth method in case of separated or nearly separated data. Based on simulations in Kessels et al. (2013), the Firth corrected multinomial logistic regression outperforms ordinary maximum likelihood as the Firth corrected method converges in case of data separation, it removes bias

of the ML estimates and reduces variance.

2.A.2 Fixed effects logistic regression: Asymptotic variance

To simplify notation we denote the vector $(\mathbf{L}, I(C=1), \ldots, I(C=m))$ by $\mathbf{X}$. To assess the uncertainty on the potential full population risk $\hat{E}\{Y(c)\}$, note that the MLE $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\psi}')'$ of $\boldsymbol{\theta}$ in

$$E(Y|\mathbf{X}) = U(\boldsymbol{\theta}, \mathbf{X}) = \text{expit}\left(\mathbf{L}'\boldsymbol{\beta} + \sum_{c=1}^{m} \psi_c \, I(C=c)\right), \tag{2.19}$$

satisfies the following set of estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \left\{Y_i - U(\hat{\boldsymbol{\theta}}, \mathbf{X}_i)\right\} = \mathbf{0}. \tag{2.20}$$

To obtain the asymptotic variance of $\hat{E}\{Y(c)\}$, we first perform a Taylor expansion of

$$\hat{E}\{Y(c)\} = \frac{1}{n}\sum_{i=1}^{n} U_c(\hat{\boldsymbol{\theta}}, \mathbf{X}_i) = \frac{1}{n}\sum_{i=1}^{n} \text{expit}\left(\mathbf{L}_i'\hat{\boldsymbol{\beta}} + \hat{\psi}_c\right) \tag{2.21}$$

around $\boldsymbol{\theta}$, which results in

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_c(\hat{\boldsymbol{\theta}}, \mathbf{X}_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_c(\boldsymbol{\theta}, \mathbf{X}_i) + E\left(\frac{\partial U_c}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{X}_i)\right)\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1). \tag{2.22}$$

Second, we perform a Taylor expansion of (2.20) around the true parameter value $\boldsymbol{\theta}$ and obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} A^{-1} R_i + o_p(1), \tag{2.23}$$

where

$$\begin{aligned} R_i &= \mathbf{X}_i \left\{Y_i - U(\hat{\boldsymbol{\theta}}, \mathbf{X}_i)\right\}, \\ A &= E\left(\frac{\partial R_i}{\partial \boldsymbol{\theta}}\right). \end{aligned}$$

Finally, we combine (2.23) and (2.22) and estimate the variance of $\hat{E}\{Y(c)\}$ via

$$\mathrm{Var}\left[\hat{E}\{Y(c)\}\right] = \frac{1}{n}\mathrm{Var}\left\{U_c(\boldsymbol{\theta},\mathbf{X}_i) + E\left(\frac{\partial U_c}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta},\mathbf{X}_i)\right)A^{-1}R_i\right\}. \tag{2.24}$$

Confidence interval calculations assume normality of logit $\hat{E}\{Y(c)\}$, so in our calculations we have transformed these estimators to the logit scale via the delta method. This allows us to obtain confidence intervals with boundaries inside the $[0,1]$ interval.

The above expressions continue to hold when the parameters $\boldsymbol{\theta}$ are estimated according to Firth's modified estimating equations (2.18). This is because bias reduction affects the covariance matrix of the estimates only in the $O(n^{-2})$ and higher order terms (Firth, 1993). Note that standard errors of the Firth corrected estimates may be slightly smaller, since they are evaluated in a different estimate for $\boldsymbol{\theta}$ (Firth, 1992).

2.A.3 Doubly robust PS method: Asymptotic variance

We denote the outcome model parameters by $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\psi}')'$ and the propensity score parameters by $\boldsymbol{\rho} = (\boldsymbol{\gamma}', \boldsymbol{\delta}')'$. We estimate $E\{Y(c)\}$ again as in (2.21), but now using a weighted regression to fit the FE model (2.19), with weights equal to one over the propensity score of the observed center $g(C_i, \mathbf{L}_i; \hat{\boldsymbol{\rho}})$, where

$$g(c,\mathbf{L}_i;\boldsymbol{\rho}) := P(C_i = c|\mathbf{L}_i) = \begin{cases} \frac{1}{1+\sum_{j=2}^{m}\exp(\mathbf{L}_i'\boldsymbol{\delta}_j+\gamma_j)} & c = 1 \\ \frac{\exp(\mathbf{L}_i'\boldsymbol{\delta}_c+\gamma_c)}{1+\sum_{j=2}^{m}\exp(\mathbf{L}_i'\boldsymbol{\delta}_j+\gamma_j)} & c \neq 1. \end{cases} \tag{2.25}$$

The MLE $\hat{\boldsymbol{\rho}}$ of $\boldsymbol{\rho}$ satisfies the following set of estimating equations

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} 1 \\ \mathbf{L}_i \end{pmatrix}\left\{I(C_i = c) - g(c,\mathbf{L}_i;\hat{\boldsymbol{\rho}})\right\} = \mathbf{0}, \qquad c = 2,\ldots,m. \tag{2.26}$$

First, we perform a Taylor expansion of (2.26) around $\boldsymbol{\rho}$ and obtain

$$\sqrt{n}(\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} A_1^{-1}R_{i1} + o_p(1), \tag{2.27}$$

where

$$\begin{aligned} R_{i1} &= \begin{pmatrix} 1 \\ \mathbf{L}_i \end{pmatrix} \left\{ I(C_i = c) - g(c, \mathbf{L}_i; \hat{\boldsymbol{\rho}}) \right\}, \\ A_1 &= E\left(\frac{\partial R_{i1}}{\partial \boldsymbol{\rho}} \right). \end{aligned}$$

Analogously, we perform a Taylor expansion of the outcome model estimating equations

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{X}_i}{g(C_i, \mathbf{L}_i; \hat{\boldsymbol{\rho}})} \left\{ Y_i - U(\hat{\boldsymbol{\theta}}, \mathbf{X}_i) \right\} = \mathbf{0} \tag{2.28}$$

around $(\boldsymbol{\theta}', \boldsymbol{\rho}')'$ and obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_2^{-1} \left\{ R_{i2} + E\left(\frac{\partial R_{i2}}{\partial \boldsymbol{\rho}} \right) \sqrt{n}(\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}) \right\} + o_p(1), \tag{2.29}$$

where

$$\begin{aligned} R_{i2} &= \frac{\mathbf{X}_i}{g(C_i, \mathbf{L}_i; \hat{\boldsymbol{\rho}})} \left\{ Y_i - U(\hat{\boldsymbol{\theta}}, \mathbf{X}_i) \right\}, \\ A_2 &= E\left(\frac{\partial R_{i2}}{\partial \boldsymbol{\theta}} \right). \end{aligned}$$

Now, let

$$U_c(\hat{\boldsymbol{\theta}}, \mathbf{X}_i) = \operatorname{expit}\left(\mathbf{L}_i' \hat{\boldsymbol{\beta}} + \hat{\psi}_c \right),$$

then we define the estimator for the potential full population risk as follows:

$$\hat{E}\{Y(c)\} = \frac{1}{n} \sum_{i=1}^{n} V_c(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\rho}}, \mathbf{X}_i) \tag{2.30}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[U_c(\hat{\boldsymbol{\theta}}, \mathbf{X}_i) + \frac{I(C_i = c)}{g(c, \mathbf{L}_i; \hat{\boldsymbol{\rho}})} \left\{ Y_i - U_c(\hat{\boldsymbol{\theta}}, \mathbf{X}_i) \right\} \right] \tag{2.31}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{I(C_i = c) Y_i}{g(c, \mathbf{L}_i; \hat{\boldsymbol{\rho}})} + \left\{ 1 - \frac{I(C_i = c)}{g(c, \mathbf{L}_i; \hat{\boldsymbol{\rho}})} \right\} U_c(\hat{\boldsymbol{\theta}}, \mathbf{X}_i) \right]. \tag{2.32}$$

Combining (2.27) and (2.29), the variance of the estimated potential full population risk $\hat{E}\{Y(c)\}$ on the risk scale is estimated via

$$\begin{aligned}\mathrm{Var}\left[\hat{E}\{Y(c)\}\right] &= \frac{1}{n}\mathrm{Var}\left[V_c(\boldsymbol{\theta},\boldsymbol{\rho},\mathbf{X}_i) + E\left(\frac{\partial V_c}{\partial \rho}(\boldsymbol{\theta},\boldsymbol{\rho},\mathbf{X}_i)\right)A_1^{-1}R_{i1}\right. \\ &\quad \left. + E\left(\frac{\partial V_c}{\partial \theta}(\boldsymbol{\theta},\boldsymbol{\rho},\mathbf{X}_i)\right)A_2^{-1}\left\{R_{i2} + E\left(\frac{\partial R_{i2}}{\partial \boldsymbol{\rho}}\right)A_1^{-1}R_{i1}\right\}\right]. \quad (2.33)\end{aligned}$$

Again, the above expressions can be transformed to the logit scale and continue to hold when the parameters $\boldsymbol{\theta}$ are estimated according to Firth's modified estimating equations (2.18).

2.B Additional Results

An illustration of model extrapolation considering two centers with strongly differential case-mix is given in Figure 2.4.

2.B.1 Simulation study application: Belgian Colorectal Cancer register

The ME models were fitted in R using the `rjags` package (Plummer, 2003) where we assigned the following independent non-informative hyperpriors:

Classical normal ME model	*Clustered normal ME model*	
$\mu_\psi \sim N(0,25)$	$\mu_k \sim N(0,25)$	$k = 1,\ldots,K$
$\sigma_\psi \sim U(0,10)$	$\sigma_k \sim U(0,10)$	$k = 1,\ldots,K$
$\beta_j \sim N(0,10^2)$	$\beta_j \sim N(0,10^2)$	$j = 1,\ldots,\ell,$

where $K = 3$ denotes the number of center clusters and ℓ is the length of the vector of patient characteristics. The Firth corrected FE model was fitted in R using the `brglm` package (Kosmidis, 2011). The analysis of 5 simulated datasets took on average 3 hours for the classical normal ME method and 15 hours for the clustered analog, as compared to at most 3 minutes for the FE and doubly robust methods.

First we determine 'true' center classifications based on a large sample of simulated data for 10^6 patients in 63 centers. Next, we generate 1000 samples

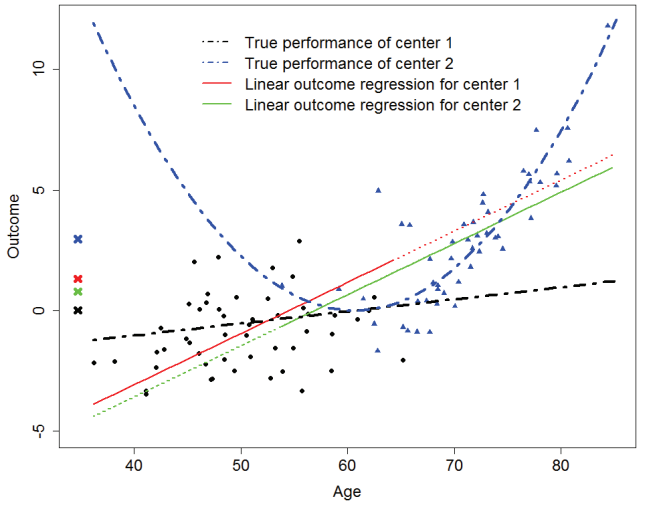


Figure 2.4: Artificial example illustrating extrapolation under outcome regression with main effects for age and center. It visualizes outcome data for two centers with strongly differential case-mix and such that smaller outcomes are seen in center 1 (black dots) than in center 2 (blue triangles) in each patient subgroup. Direct standardization under seemingly well fitting linear regression models involves strong extrapolation (dashed line) of the outcome distribution in the given center to what it is expected to achieve on the full patient population under study. In truth, the observed data carry no information about $E\{Y(2)\}$ because no patients in center 2 have age below 60.

of each 2355 patients under the same models. In this data-generating process, we first generated independent patient characteristics age, gender and initial disease status cStage based on descriptive statistics in Goetghebeur et al. (2011), such that age was 67 years on average (SD 9 years), that 61% of the patients were male and that 13%, 18%, 51%, 14% respectively had cStage I to IV; 4% had missing cStage. Next, we use the PS and outcome models that were fitted on the observed data as data-generating models. So, the patients' center choice and outcome are generated, mimicking the center-specific distribution of patients in the register. Then, based on the large random sample of 10^6 patients we obtain precise estimates of $E(Y)$ and $E\{Y(c)\}$ as sample averages. Following equations (2.2) and (2.3) in the main text, this results in the true center classification. For each center this true classification is then compared to the classification based on a sample of $n = 2355$ patients and using one of the regression methods to estimate $\hat{E}(Y)$ and $\hat{E}\{Y(c)\}$. In the analysis, we excluded centers treating less than 10 patients, resulting in a median of 33 registered patients per center and an average of 53 centers.

About half of the centers are labelled as having low/high risk based on the initial large sample (see main text Figure 1, ignoring the simulation results for now) although for many centers the 'true' potential full population risk is close to the decision limits, especially on the low side. Note that results depend highly on the values of the clinical (λ) and statistical (k) tolerance levels. When early detection is most important, such as for confidential feedback to clinical centers, large power is preferred over small type I errors. In contrast, when publishing results openly, mistakenly classifying a center as high risk may have unduly severe implications for that center (decrease in number of patients, lower funding) compared to mistakenly failing to recognize a center as potentially high risk.

Tables 2.3 and 2.4 provide extra results on simulations modelled on the Belgian Colorectal Cancer register when the Firth correction is not applied and when the sample size is doubled. We performed additional simulations to investigate the effect of model misspecification. We generate data under an outcome regression model with quadratic centered age effect and a PS model with the absolute value of centered age effect. Misspecification is introduced in the fitted outcome and/or PS model by estimating a linear centered age effect instead. The

	Fixed effects	Doubly robust PS
Power (%)		
to detect High	59.6	52.6
to detect Low	38.5	45.3
Type I error (%)		
Low as High	1.8	5.5
Accepted as High	7.4	10.5
Type I error (%)		
High as Low	1.5	4.2
Accepted as Low	14.9	23.9
Coverage of 95% *CI*		
for $E\{Y(c)\}$ *(%)*	94.4	88.5
Classification		
% High (22%)	15.3	16.9
% Low (30%)	19.6	25.9
% correct (L-A-H)	62.5	57.6

Table 2.3: Center classification (Low/High risk or Accepted) based on 1000 *simulations for FE and doubly robust PS method without Firth correction. The true percentage of low and high risk centers is respectively* 30% *and* 22%.

original population average risk $E(Y) = 23\%$ based on Goetghebeur et al. (2011) is retained. Center effects were multiplied by 2 to obtain a similar spread of $E\{Y(c)\}$ values compared to the original simulations. The original age coefficients in the PS models are multiplied by 2 to obtain more extreme differences in age distribution and thus stronger extrapolation across centers.

Descriptives under the new and original data generating models for a random sample of $n = 2355$ patients are respectively given in Figures 2.5 and 2.6, where centers with size smaller than 10 are indicated in red. The quadratic age effect is quite pronounced while the corresponding estimated linear age effect is approximately zero (See bottom of Figure 2.5). Comparing the age distribution under both simulation scenarios, we see more overlap in the age distribution across centers under the new data generating PS model.

Simulation results in table 2.5 show that the Firth corrected fixed effects and doubly robust PS methods have similar performance when the regression models are correctly specified. We summarize the bias on $E\{Y(c)\}$ over all centers by taking the difference of the mean of the squared bias per center and the mean of the empirical variance of the bias over simulations and next taking the square

	Classical normal ME	Clustered normal ME	Firth corrected FE	Firth corrected doubly robust PS
Power (%)				
to detect High	41.2	40.3	60.6	55.1
to detect Low	22.8	13.4	36.1	39.7
Type I error (%)				
Low as High	0.1	0.2	1.2	3.5
Accepted as High	1.0	1.8	6.4	9.4
Type I error (%)				
High as Low	0.2	0.2	0.6	1.8
Accepted as Low	4.1	4.4	10.1	16.6
Coverage of 95% *CI for* $E\{Y(c)\}$ *(%)*	95.8	84.4	94.8	90.6
Classification				
% High (22%)	9.5	9.8	16.9	17.7
% Low (30%)	9.2	6.5	16.7	21.2
% correct (L-A-H)	62.7	59.0	66.6	61.8

Table 2.4: Center classification (Low/High risk or Accepted) based on 1000 *simulations, except for the Bayesian normal ME analyses which were evaluated on* 100 *simulations,* ***with double sample size****. The true percentage of low and high risk centers is respectively* 30% *and* 22%.

root, that is

$$\sqrt{E_c\left[\text{Bias}(E\{Y(c)\})^2\right] - E_c\left[\frac{\text{Var}\left(\hat{E}\{Y(c)\}\right)}{S}\right]}, \tag{2.34}$$

where S is the number of simulations and E_c is the average over all centers $c = 1, \ldots, m$. Note that we correct for the fact that the squared bias also depends on the variance of the bias. In general the coverage of the 95% CI for $E\{Y(c)\}$ is smaller than 95%, but this may be due to extrapolation across centers, which is also reflected in the bias. For both methods the power to detect high mortality risk centers decreases when the outcome model is misspecified, but the Type I errors do not decrease accordingly. Here, the doubly robust PS method outperforms the fixed effects method in terms of coverage of the 95% CI for $E\{Y(c)\}$, percentage of correct classification and bias on $E\{Y(c)\}$, although it does not gain power to detect high mortality risk centers. Misspecification of the PS model does not seem to have a large impact on the results in our simulation study, although this may be different for other data generating models. When both models are

	Both models correct		Misspecification of Outcome model		PS model	Both models
	FE	DR	FE	DR	DR	DR
Power (%)						
to detect High	65.4	64.7	54.2	50.1	64.8	50.2
to detect Low	41.6	43.0	40.1	41.8	40.4	47.4
Type I error (%)						
Low as High	1.5	4.5	5.3	5.0	4.7	7.0
Accepted as High	5.9	11.0	10.5	10.0	11.6	14.2
Type I error (%)						
High as Low	0.7	0.8	1.6	1.6	0.6	5.0
Accepted as Low	11.3	13.5	20.2	14.9	13.6	26.8
Classification						
% High (19%)	14.8	15.8	18.3	13.4	15.9	18.8
% Low (19%)	15.9	17.7	20.7	16.8	17.7	25.7
% correct (L-A-H)	72.9	70.1	60.3	66.2	70.0	56.8
Bias on $E\{Y(c)\}$ $(\times 10^{-2})$	1.54	3.62	5.85	5.63	3.52	8.32
MSE of $E\{Y(c)\}$ $(\times 10^{-2})$	0.330	0.447	0.922	1.01	0.458	1.32
For 95% *CI of* $E\{Y(c)\}$						
coverage (%)	88.7	83.2	84.2	90.2	84.2	78.6
median length	0.167	0.167	0.288	0.323	0.167	0.299

Table 2.5: Center classification (Low/High risk or Accepted) based on 1000 *simulations, under model misspecification. FE = Fixed effects method, DR = Doubly robust PS method, both with Firth correction. The true percentage of low and high risk centers is* 19%.

misspecified, the doubly robust PS method performs worse than the fixed effects method with misspecified outcome model.

2.B.2 Analysis of the Swedish Stroke Register

We give descriptive statistics for both the original and reduced dataset in Table 2.6. To assess the extent of differential patient case-mix across centers, we perform a univariate analysis on the complete cases (Figures 2.9 to 2.11). In general case-mix does not differ much across centers, but we detect considerable variation with respect to treatment for high blood pressure (Figure 2.7), education level (Figure 2.8) and time of admission (Figure 2.9). The minimum number of patients for whom we have records per hospital is 104 (median 1383.5 and max 7260) for the full data and 49 (median 957.5 and max 4341) for the complete cases. Some

centers have no records for several years because they no longer treat stroke patients, have been merged with another center, or only started treating stroke patients during the study. There are no missing values for the outcome and the 30-day mortality risk is lower for the complete cases (8%) than for the full data (12%), which is due to the complete cases being generally more healthy as can be seen from the patient characteristics measured on admission (see Table 2.6). Figure 2.11 shows lower center-specific mortality risks for the complete cases compared to all cases, so results will probably be too optimistic for all centers. To describe the effect of risk factors on outcome, we report the estimated odds ratios based on fitting a logistic normal ME model (see Table 2.7).

The analysis is based on a logistic model for the outcome in which we allowed for a piecewise linear spline effect of standardized age (mean age 75 years, SD = 12 years) with a knot at age 80. Besides the main effects listed in Table 2.7, we include three interactions between patient-specific covariates which are of primary interest: education level by age exceeding 80 to account for a different effect of education level on 30-day mortality for elderly or young patients, institution by living alone to account for a different effect of living alone for patients in an institution or at home, and living alone by gender to account for a different effect of living alone for men and women. Note that we do not include interactions with center in the model. The time-dependent PS model includes the same risk factors, with exception of the interactions, and it does not allow for a piecewise linear effect of age to avoid model fitting problems. Results for the analysis of the Swedish Stroke Register when applying the Firth correction to the outcome model for the FE method and the doubly robust method are given in Figures 2.12 to 2.14.

Risk Factor	Full data Prevalence (%)	Missing (%)	Complete cases Prevalence (%)
Male	50.3	0	53.5
p-ADL dependence (*)	7.2	1.3	4.9
Institutional living (*)	7.0	0.5	4.0
Living alone (*)	48.9	0.7	42.7
Atrial fibrilation (*)	25.3	1.2	21.4
Diabetes (*)	18.9	0.5	19.1
Trt for high blood pressure (*)	50.3	1.4	50.3
Current smoker (*)	16.2	12.8	18.8
Stroke subtype		0	
(Intracerebral haemorrhage (I61))	12.1		11.8
Cerebral infarction (I63)	83.6		84.8
Unspecified stroke (I64)	4.3		3.4
Consciousness at admission		0.8	
(Alert)	82.9		86.8
Drowsy	12.1		9.7
Unconscious	5.0		3.5
Education		20.8	
(Primary)	51.9		51.6
Secondary	34.3		34.5
University	13.7		13.8
Adjusted yearly income (in 100 SEK)		0.5	
(< 185)	10.0		9.4
815 to 1235	39.0		34.0
1235 to 2241	40.4		44.1
> 2241	10.6		12.5
Year of admission		0	
(2001)	9.5		7.4
2002	10.3		8.7
2003	10.8		9.4
2004	11.1		10.6
2005	11.8		11.8
2006	11.6		12.0
2007	11.4		12.2
2008	11.7		13.0
2009	11.8		14.9

Table 2.6: Descriptive percentages of the binary and categorical covariates from the Swedish Stroke Register, indicating (*) if measured prior to stroke and (reference category) for analysis. For the full data we calculated percentages for each risk factor not taking into account missing values for that factor.

Fixed effects	$\hat{\beta}$	SE($\hat{\beta}$)	$\widehat{\text{OR}}$	95% CI for OR	p-value
(Intercept)	−2.854	0.084	0.058	0.049 to 0.068	< 0.001
Male	0.163	0.040	1.178	1.088 to 1.275	< 0.001
Age (standardized)	0.513	0.024	1.671	1.594 to 1.751	< 0.001
Age > 80 years	0.441	0.089	1.555	1.306 to 1.851	< 0.001
p-ADL dependence (*)	0.625	0.050	1.869	1.693 to 2.063	< 0.001
Institutional living (*)	0.533	0.123	1.703	1.337 to 2.169	< 0.001
Living alone (*)	0.020	0.043	1.020	0.938 to 1.109	0.639
Atrial fibrilation (*)	0.444	0.030	1.559	1.469 to 1.654	< 0.001
Diabetes (*)	0.205	0.034	1.227	1.149 to 1.311	< 0.001
Trt for high blood pressure (*)	−0.026	0.028	0.974	0.922 to 1.030	0.355
Current smoker (*)	0.114	0.040	1.121	1.035 to 1.213	0.005
Stroke subtype					
(vs Intracerebral haemorrhage (I61))					
Cerebral infarction (I63)	−0.933	0.034	0.393	0.368 to 0.421	< 0.001
Unspecified stroke (I64)	−0.419	0.072	0.657	0.571 to 0.757	< 0.001
Consciousness at admission (vs Alert)					
Drowsy	1.885	0.031	6.589	6.202 to 7.001	< 0.001
Unconscious	3.445	0.043	31.340	28.827 to 34.072	< 0.001
Education (vs Primary)					
Secondary	−0.094	0.035	0.910	0.849 to 0.976	0.008
University	−0.146	0.053	0.864	0.779 to 0.958	0.005
Adjusted yearly income (in 100 SEK)					
(vs < 185)					
815 to 1235	0.043	0.049	1.044	0.948 to 1.149	0.381
1235 to 2241	−0.012	0.049	0.988	0.898 to 1.088	0.811
> 2241	−0.146	0.065	0.864	0.760 to 0.982	0.025
Year of admission (vs 2001)					
2002	−0.071	0.074	0.931	0.806 to 1.076	0.335
2003	−0.025	0.071	0.975	0.848 to 1.122	0.726
2004	0.081	0.069	1.085	0.948 to 1.241	0.238
2005	0.115	0.068	1.122	0.983 to 1.281	0.089
2006	0.108	0.067	1.114	0.976 to 1.272	0.108
2007	0.104	0.067	1.109	0.972 to 1.266	0.123
2008	0.091	0.067	1.095	0.960 to 1.249	0.177
2009	0.097	0.067	1.101	0.965 to 1.257	0.151
Male × Living alone	−0.022	0.057	0.978	0.875 to 1.094	0.701
Institutional living × Living alone	0.030	0.131	1.031	0.797 to 1.333	0.816
Age > 80 years ×					
Education (Second. vs Primary)	0.050	0.135	1.051	0.807 to 1.368	0.711
Education (Univ. vs Primary)	−0.036	0.213	0.965	0.635 to 1.466	0.867
Random center effects ($\widehat{\text{Var}}$)	0.069				

Table 2.7: Model parameters for a descriptive logistic normal ME model with outcome 30-day mortality based on the Swedish Stroke Register, indicating (*) if measured prior to stroke and (vs reference category) for categorical covariates. SE = standard error, OR = odds ratio = $\exp(\beta)$, 95% CI for OR = 95% confidence interval for odds ratio.

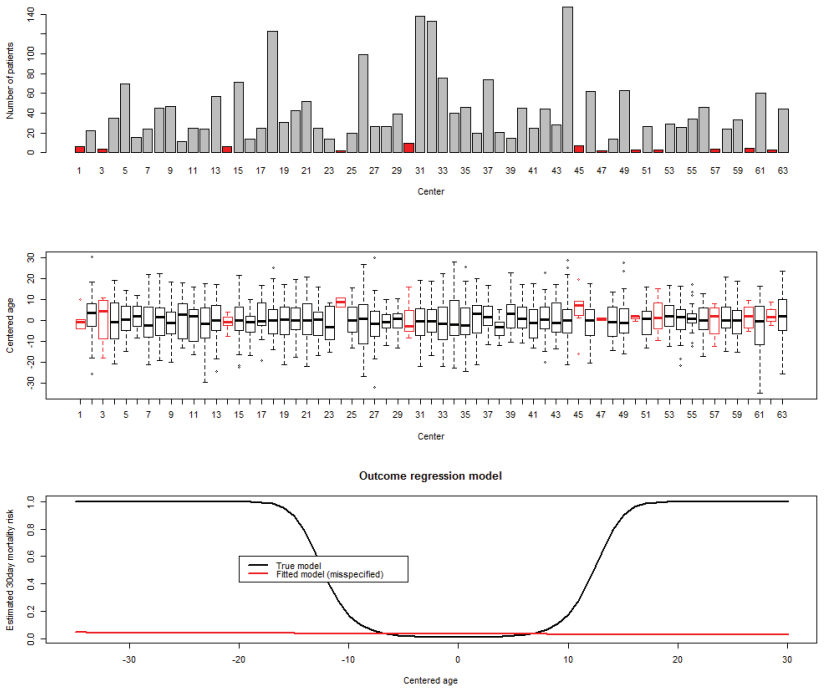


Figure 2.5: Age distribution under the new data generating PS model (model misspecification), for a random sample of n = 2355 patients. Figure at the bottom: Fitted misspecified outcome model is a Firth corrected fixed effects model and estimated risks are for female patients with cStageI.

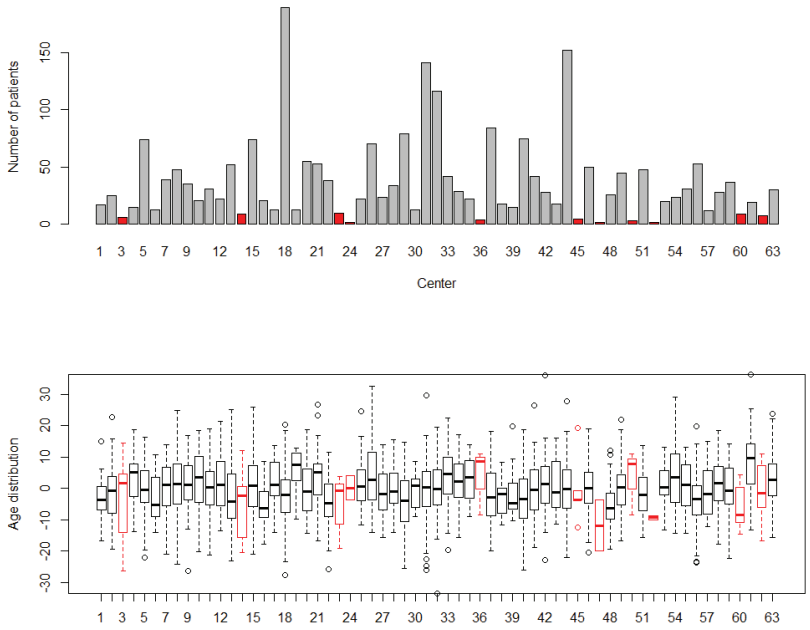


Figure 2.6: Age distribution under the data generating PS model based on the Belgian colorectal cancer register.

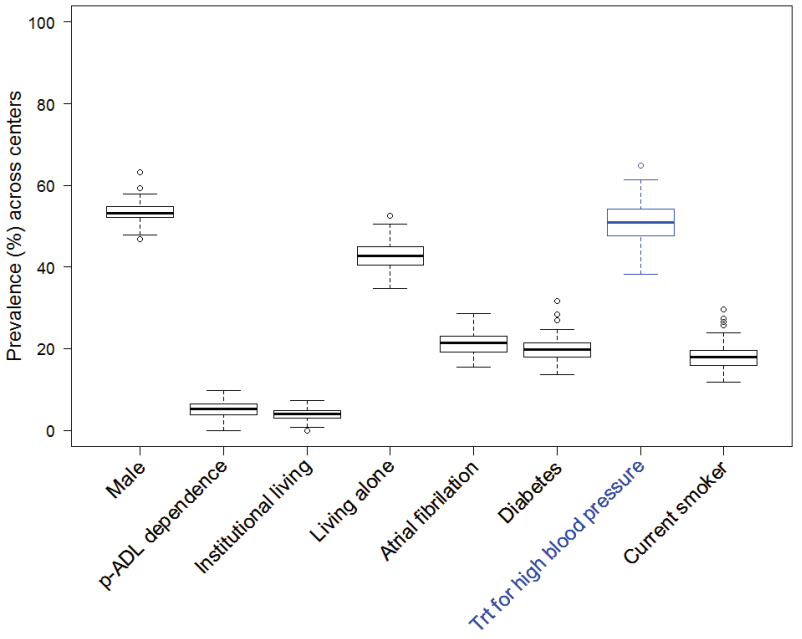


Figure 2.7: Prevalence of the binary patient-specific covariates across centers based on the complete cases in the Swedish Stroke Register.

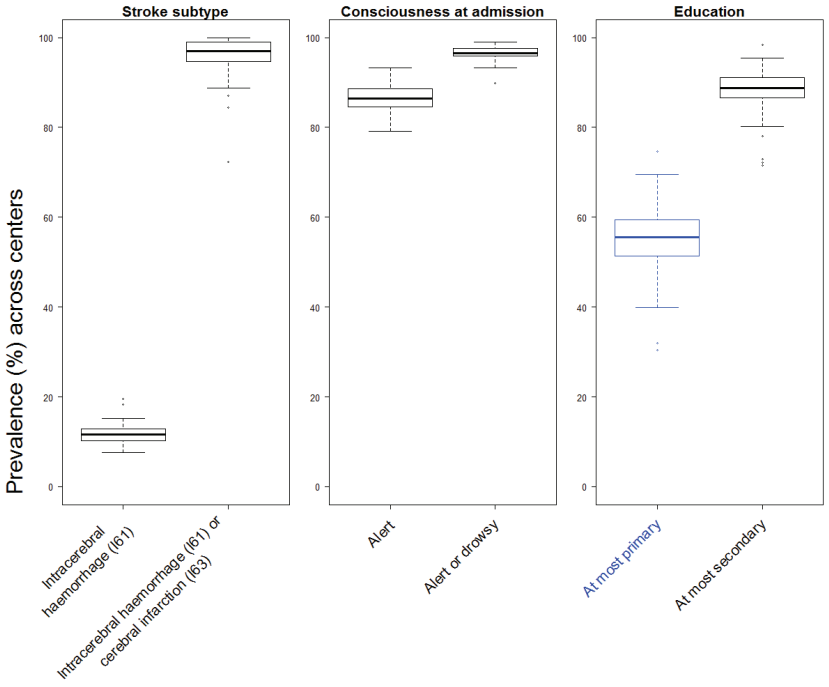


Figure 2.8: Prevalence of stroke subtype, consciousness at admission and education across centers based on the complete cases in the Swedish Stroke Register.

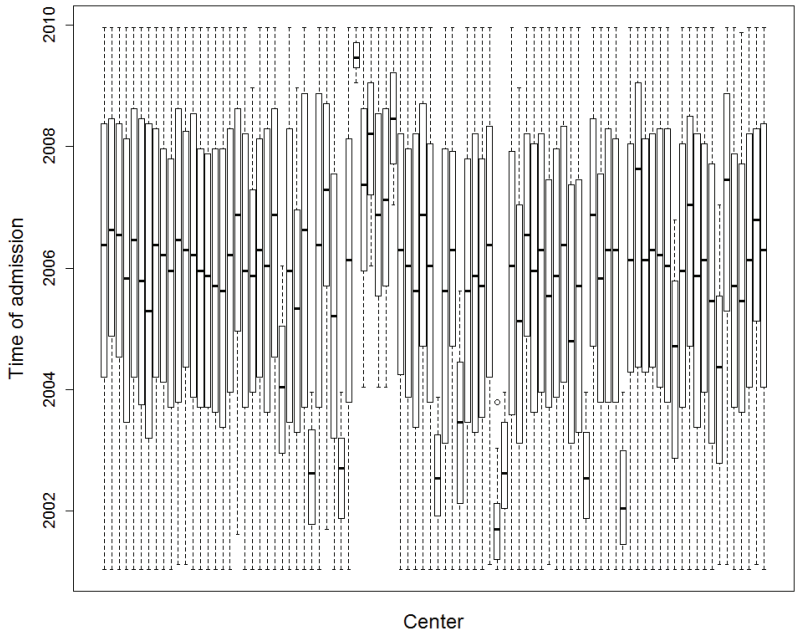


Figure 2.9: Boxplot for the distribution of time of patient admission across centers based on the complete cases in the Swedish Stroke Register.

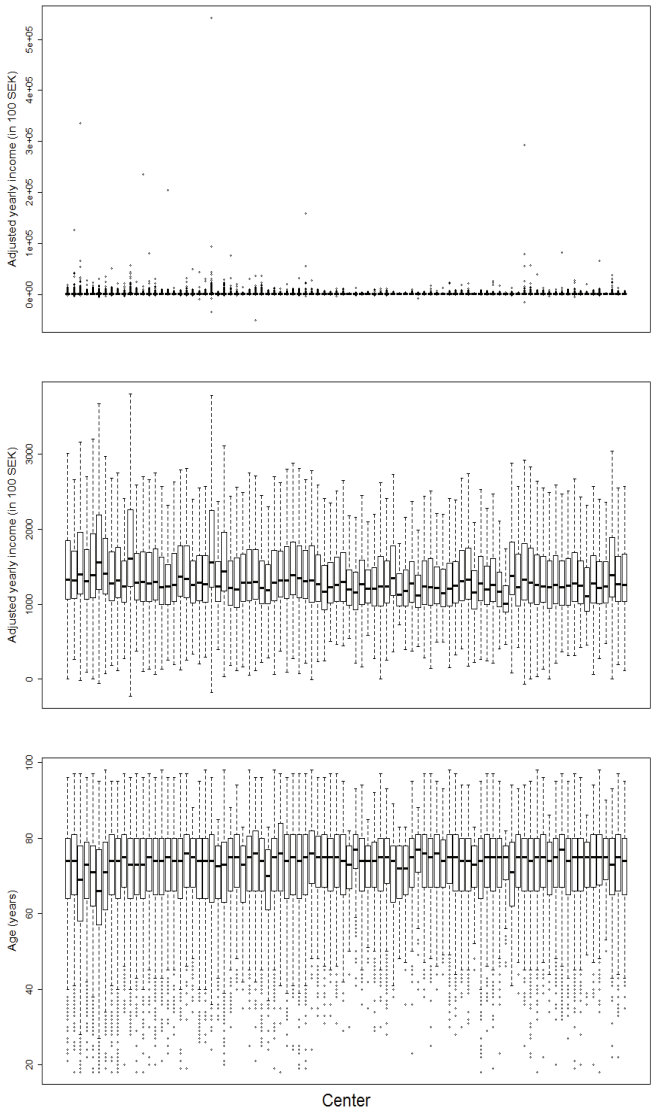


Figure 2.10: Boxplots for the distribution of adjusted yearly income (in 100 SEK) with-/without outliers and age per center based on the complete cases in the Swedish Stroke Register.

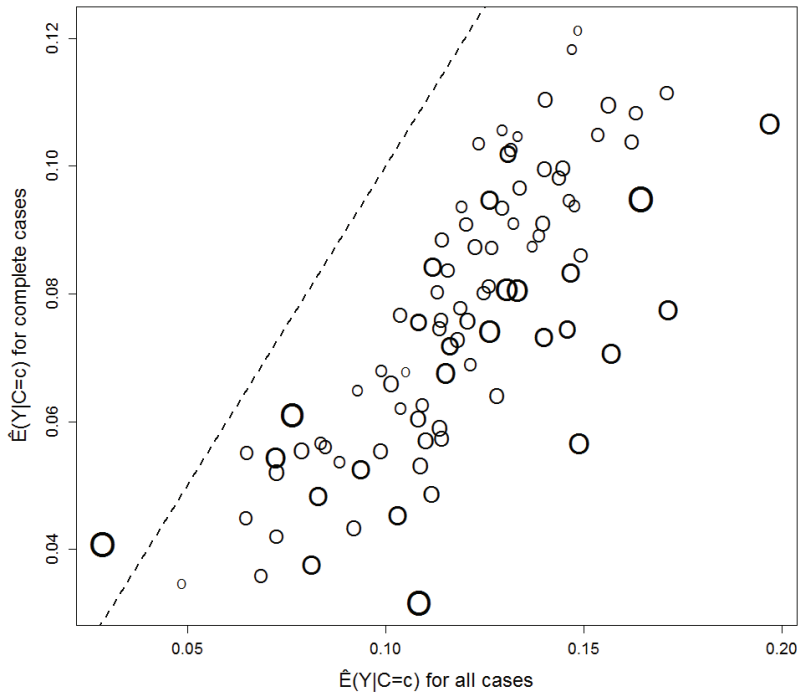


Figure 2.11: Observed mortality risk per center for all cases and complete cases in the Swedish Stroke Register. The dashed line represents the first bisector and bulb sizes are proportional to the percentage of patients with missing values in that center.

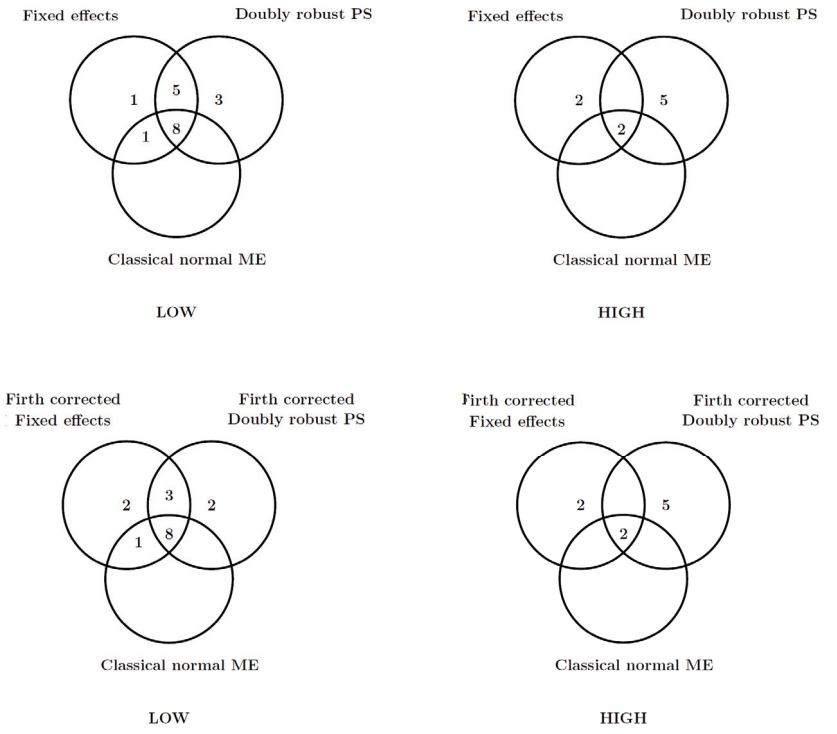


Figure 2.12: Center classification for the 90 centers in the Swedish Stroke Register, without and with Firth correction.

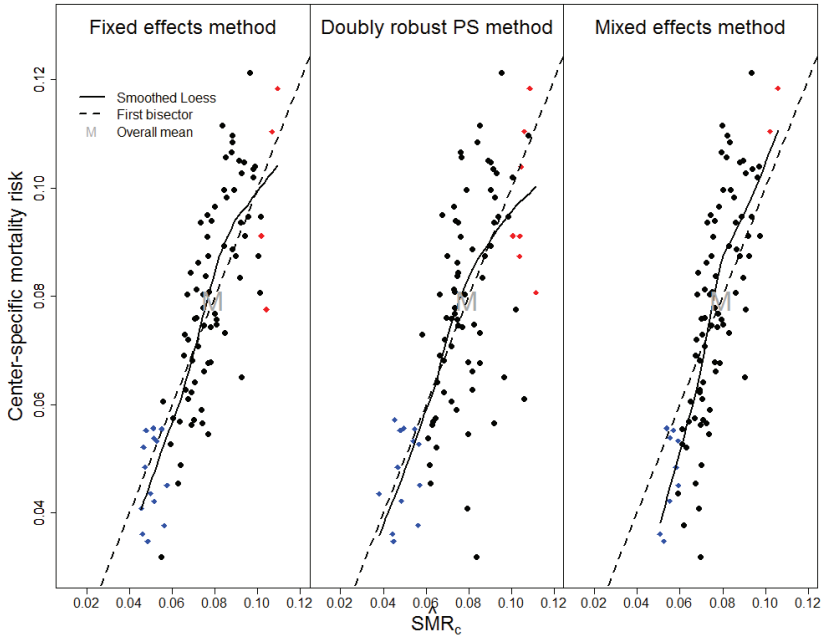


Figure 2.13: The observed center-specific risk versus the potential full population risk for all 90 centers of Riksstroke for the FE method, the doubly robust PS method and the classical normal ME method, distinguishing between low (blue filled circles) and high (red triangles) mortality risk (with Firth correction).

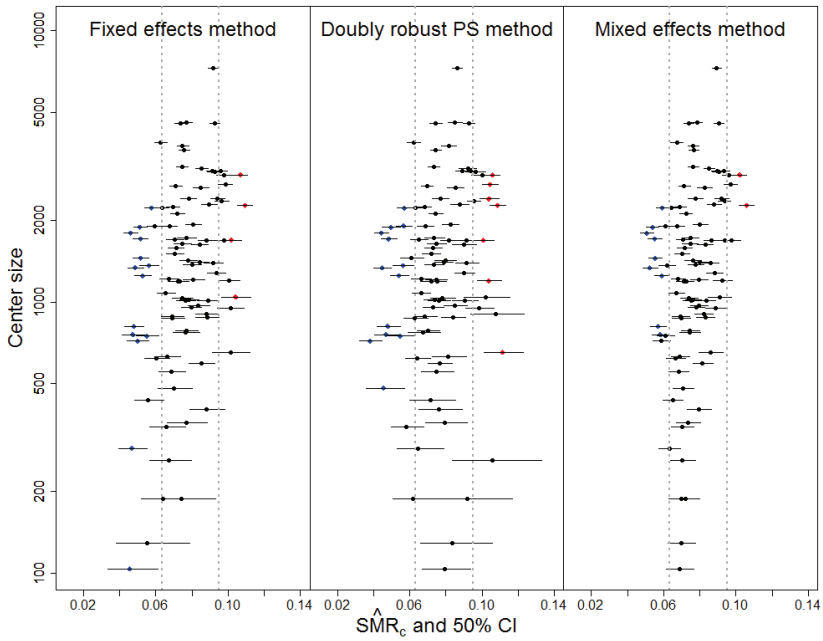


Figure 2.14: The potential full population risk and corresponding 50% confidence interval for all 90 centers of Riksstroke for the FE method, the doubly robust PS method and the classical normal ME method. The vertical gray lines indicate the clinical decision limits $(1-\lambda)\hat{E}(Y)$ *and* $(1+\lambda)\hat{E}(Y)$*, with* $\lambda = 0.20$*, which are used for classification of low (blue filled circles) and high (red filled circles) mortality risk centers (with Firth correction).*

CHAPTER 3

On the Practice of Ignoring Center-Patient Interactions in Evaluating Hospital Performance

This chapter is based on the following paper: Varewyck, M., Vansteelandt, S., Eriksson, M., and Goetghebeur, E. (2015). "On the practice of ignoring center-patient interactions in evaluating hospital performance," *Statistics in Medicine*.

Summary

We evaluate the performance of medical centers based on a continuous or binary patient outcome (e.g. 30-day mortality). Common practice adjusts for differences in patient mix through outcome regression models which include patient-specific baseline covariates (e.g. age, disease stage) besides center effects. Since a large number of centers may need to be evaluated, the typical model postulates that the effect of center on outcome is constant over patient characteristics. This may be violated, for example when some centers are specialized in children or geriatric patients. Including interactions between certain patient characteristics and the many fixed center effects in the model increases the risk for overfitting however, and could imply a loss of power for detecting centers

with deviating mortality. Therefore, we assess how the common practice of ignoring such interactions impacts the bias and precision of directly and indirectly standardized risks. The reassuring conclusion is that the common practice of working with main effects of center has minor impact on hospital evaluation, unless some centers actually perform substantially better on a specific group of patients and there is strong confounding through the corresponding patient characteristic. The bias is then driven by an interplay of the relative center size, the overlap between covariate distributions and the magnitude of the interaction effect. Interestingly, the bias on indirectly standardized risks is smaller than on directly standardized risks. We illustrate our findings by simulation and in an analysis of 30-day mortality on Riksstroke.

3.1 Introduction

Many continuing efforts are made to improve the accuracy of hospital quality of care assessments (Normand et al., 1997; Shahian et al., 2004). They are motivated by the major impact of performance evaluations not only on the improvement of care, but also on the patient's choice of hospital, or financial pay-per-performance incentives for example. Key aspects of the quality of hospital performance are commonly evaluated through a binary or continuous outcome quality indicator via direct or indirect standardization (Shahian and Normand, 2008; Spiegelhalter, 2005a). Direct standardization aims to assess for each center how the entire study population would have fared under its current level of care. Indirect standardization contrasts the quality outcome in each center with what is expected should their patients choose randomly over the level of care across all centers. Since the choice of standardization technique depends on the research question, we will report results for both techniques. Traditionally indirect standardization is used and this is most relevant to judge center performance when the center's own population does not substantially change over time. Should centers vary in approach and one wishes to choose one approach for implementation across all centers, then direct standardization delivers the most relevant impact measure.

In general, standardized risks are obtained via outcome regression models that adjust for confounding of the center-outcome effect (DeLong et al., 1997; He et al., 2013). Adjustment for differential patient mix is necessary, since centers treating for instance older patients will show a higher mortality risk irrespective of their actual treatment quality. Interactions between center and patient characteristics are rarely modelled, however. This is due to the curse of dimensionality which may already show up in the main effects model. One may hit low information content to estimate the main effects of many centers, resulting in large finite sample bias. For this reason we have chosen to use Firth corrected fixed effects regression instead of standard fixed effects regression (Varewyck et al., 2014). This method is preferred over standard random effects regression which may unduly shrink center effects towards the overall mean (Normand et al., 1997), thereby masking outlying performance of especially the smallest centers (Varewyck et al., 2014).

In practice, center effects may interact with patient characteristics when some centers perform particularly well on a specific subgroup of patients (Shahian et al., 2004; Gatsonis et al., 1993, 1995). In Austin et al. (2003) for instance, differences in estimated 30-day mortality risk between hospitals were relatively small for patients with low illness severity but variation increased for patients with increasing illness severity. This may indicate that high-risk patients receive more specialized care at some hospitals. Similarly in Normand et al. (1997), the variation in hospital effect was substantial and depended on patient baseline severity. In a next step it is important to know which hospitals and why they show interactions. Evidence suggested that effect differences were sometimes associated with hospital size, urbanicity and academic affiliation, as medium sized hospitals had slightly weaker effects of baseline severity on the outcome than large hospitals.

In this article, we will therefore study to what extent bias may enter each of the standardized risks when modelling constant center effects across patient profiles while interactions between center and patient characteristics are present. This may justify the common practice of ignoring center-patient interactions, especially in situations where it is simply prohibitive to allow for effect modification because sufficient information is lacking in small centers, for example. In

Saposnik et al. (2007) fitting problems are overcome by modelling interactions between patient characteristics and hospital type (teaching status, location), but this limits evaluations to hospital groups rather than individual hospitals.

3.2 Setting

Throughout the paper, C will denote a random variable indicating in which center the patient was actually treated ($C = 1, \ldots, m$) and $\mathbf{L}$ denotes the vector of patient-specific baseline characteristics such as gender, age and initial disease status. We focus on the following data-generating outcome regression model:

$$E(Y|\mathbf{L}, C) = g\left(\sum_{c=1}^{m} I(C = c)(\mathbf{L}'\beta_c + \psi_c)\right), \tag{3.1}$$

which allows center effects to depend on $\mathbf{L}$. So the expected outcome in center c is parametrized by ψ_c for a patient with the reference ($\mathbf{L} = \mathbf{0}$) profile, and by $\psi_c + \mathbf{l}_0'\beta_c$ for a patient with $\mathbf{L} = \mathbf{l}_0$ profile. Here, $g(\cdot)$ is a known link function, e.g. the logistic link.

3.2.1 Nature of Interactions

In the outcome regression model (3.1), interactions between center and patient-specific characteristics may arise in different ways. We are interested in the case where some centers perform structurally better on a specific subgroup. For example, the difference in care between hospitals is not constant among age groups when younger patients get very similar care in each center while older patients receive much better care in some centers compared to others, perhaps due to special equipment or experience of the hospital staff with geriatric patients.

In the absence of such structural interactions, center-patient interactions could still occur due to the scale of the fitted model. While center effects may not interact with patient characteristics on the scale of the linear predictor in a logistic regression model, an interaction may be needed if an additive linear model is used instead (Greenland et al., 1999). Interactions may also manifest

themselves as a result of unmeasured confounding e.g. due to unknown environmental factors. For example, pollution may increase mortality risk in some regions. If the performance of each center is constant over age but the pollution especially affects older patients, then it will induce poorer center effects for older patients in polluted regions. Unmeasured confounders may thus introduce or hide interactions between center and measured confounders (VanderWeele et al., 2012). Throughout, we will exclude this possibility as we will assume that there are no unmeasured confounders, i.e.

$$Y(c) \perp\!\!\!\perp C|\mathbf{L} \text{ for all } c, \tag{3.2}$$

where $Y(c)$ indicates the potential outcome for given patient if he/she were treated at the care level of center c (Hernán and Robins, 2006b).

3.2.2 Direct and Indirect Standardization

We will assess the impact on the directly or indirectly standardized risk of the practice of ignoring interactions between center and patient characteristics in evaluating center performance.

Direct standardization aims to infer the potential full population risk for each center c: the risk that would be realized if all patients under study were to experience the care level of that given center c, irrespective of where they were actually treated. We denote this by $E\{Y(c)\}$. Under the outcome regression model in (3.1) and assuming (3.2), this can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{L}_i'\hat{\boldsymbol{\beta}}_c + \hat{\psi}_c), \tag{3.3}$$

for a study population of size n, where $\hat{\boldsymbol{\beta}}_c$ and $\hat{\psi}_c$ are Firth penalized-likelihood estimators (Firth, 1993). We can then make pairwise comparisons between the directly standardized risk of different centers or with the overall mortality risk $E(Y)$ estimated by

$$\frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{3.4}$$

In contrast, indirect standardization focuses on what a center achieves for its own patient mix. In general, a risk ratio or risk difference is measured between the observed and expected (e.g. averaged over all care levels) risk in each center (Shahian et al., 2001). For instance, the excess risk takes the difference between the center's observed risk and the expected risk if its patients were randomly assigned to the care level across the observed distribution of centers, i.e.

$$\text{Excess risk} = E\{Y(c)|C=c\} - \frac{1}{m}\sum_{c^*=1}^{m} E\{Y(c^*)|C=c\}. \tag{3.5}$$

Here, the observed risk in center c, $E\{Y(c)|C=c\}$, is estimated by

$$\frac{\sum_{i=1}^{n} Y_i I(C_i=c)}{\sum_{i=1}^{n} I(C_i=c)}. \tag{3.6}$$

Under the outcome regression model in (3.1) and assuming (3.2), the expected risk under the average care level for patients of center c is estimated as:

$$\frac{\sum_{i=1}^{n} m^{-1}\sum_{c^*=1}^{m} g\left(\mathbf{L}_i'\hat{\boldsymbol{\beta}}_{c^*} + \hat{\psi}_{c^*}\right) I(C_i=c)}{\sum_{i=1}^{n} I(C_i=c)}. \tag{3.7}$$

3.2.3 Ignoring interactions

Below, we will evaluate the bias on estimators (3.3) and (3.7) when the working outcome model involves a common center effect γ_c over all patient profiles instead of covariate-specific center effects:

$$E(Y|\mathbf{L},C) = g\left(\mathbf{L}'\boldsymbol{\beta} + \sum_{c=1}^{m} I(C=c)\gamma_c\right). \tag{3.8}$$

Here, the effect of center c on patient's outcome is expressed by the parameter γ_c which is now assumed to be the same for each given patient profile.

3.3 Asymptotic Bias Calculation

We calculate the asymptotic bias on the directly and indirectly standardized risk when imposing a constant center effect among patients instead of allowing for

center-patient interactions in (3.3) and (3.5). The average of the observed risks in center c is not model-based when estimated by (3.6) and therefore unbiased. So we will calculate the bias on the indirectly standardized risk for center c through the bias on the expected risk when care levels are averaged over all centers, $m^{-1}\sum_{c^*} E\{Y(c^*)|C=c\}$. For simplicity, we focus first on linear regression models including m centers and one patient characteristic L. We fix the number of centers m in our asymptotic calculations, since we focus on the evaluation of centers in a setting where m is relatively fixed (e.g. Riksstroke), but patients come and go. In the Appendix (Section 3.A) we provide details on the calculations, which are based on a similar principle as in (Liu and Gustafson, 2008).

The asymptotic bias on the directly standardized risk in center c is given by:

$$\{E(L|C=c)-E(L)\}\left[\beta_c-\sum_{j=1}^{m}\frac{P(C=j)\mathrm{Var}(L|C=j)}{E\{\mathrm{Var}(L|C)\}}\,\beta_j\right]. \tag{3.9}$$

For the asymptotic bias on the excess risk for center c we obtain:

$$m^{-1}\sum_{c^*=1}^{m}\left\{E(L|C=c^*)-E(L|C=c)\right\}\left[\beta_{c^*}-\sum_{j=1}^{m}\frac{P(C=j)\mathrm{Var}(L|C=j)}{E\{\mathrm{Var}(L|C)\}}\,\beta_j\right]. \tag{3.10}$$

The first factor in these expressions refers to the difference in patient mix between centers, while the second factor contrasts a center-specific L-effect with a weighted average of interaction effects.

Starting with the first factor, it is obvious that for both standardized risks the bias is zero when all centers have exactly the same L-distribution, so in particular when L is no confounder of the center-outcome effect; in fact it suffices that the mean of L is equal in all centers. If not, strong confounding by L implies a small overlap in patient mix between centers or a large 'extrapolation distance' of results from one center to the other, and thus may lead to large bias. The bias increases for a larger deviation of the mean of L in center c from either the mean in the overall population for direct standardization (3.9) or the mean in any other center for indirect standardization (3.10). This difference between both standardization techniques can be explained by different extrapolation, and is illustrated in Figure 3.1 for two centers. Direct standardization extrapolates the

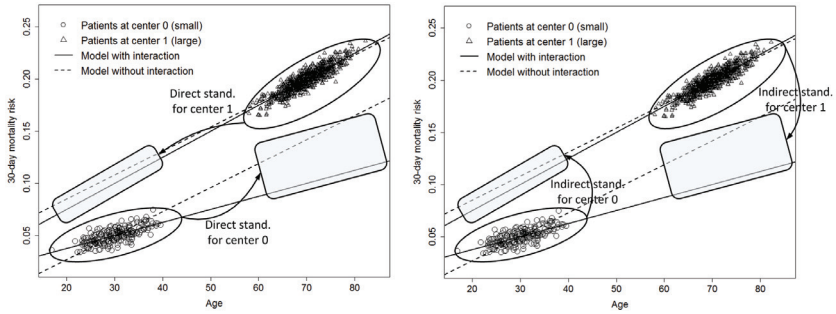


Figure 3.1: Extrapolation in the directly and indirectly standardized risk considering 2 centers (small or large center size). The 30-day mortality risk is estimated by a model with or without interaction between center and patient's age.

estimated performance at center c to the whole population under study, while for indirect standardization the performance of any other center is extrapolated to the patients in center c.

It is clear that the second factor, and thus the bias, is zero when there are no center-patient interactions because then $\beta_1 = \ldots = \beta_m$. We recognize the same weighted sum of interaction effects for both standardization techniques. For a center j the weight $P(C = j)\text{Var}(L|C = j)E\{\text{Var}(L|C)\}^{-1}$ corresponds to the relative spread of the patient mix in that center compared to the other centers, where a larger center size or larger variance of the center-specific L-distribution will result in a larger weight. The weighted sums are then respectively compared to the interaction effect in center c for direct standardization or the interaction effect in any other center than c for indirect standardization. However, in both cases stronger interactions will result in larger bias.

To get more insight in the difference between the bias for direct and indirect standardization, we consider $m = 2$ centers, now coded as $c = 0$ and $c = 1$. Then, the bias on the directly standardized risk for center c (3.9) simplifies to

$$\{E(L|C = c) - E(L)\}\,(\beta_c - \beta_{1-c})P(C = 1-c)\,\text{Var}(L|C = 1-c)\,E\{\text{Var}(L|C)\}^{-1}, \quad (3.11)$$

and for indirect standardization (3.10) to

$$\frac{1}{2}\{E(L|C=1-c)-E(L|C=c)\}(\beta_{1-c}-\beta_c)P(C=c)\,\mathrm{Var}(L|C=c)\,E\{\mathrm{Var}(L|C)\}^{-1}. \tag{3.12}$$

For both standardizations we again recognize how the difference in average covariate levels and the magnitude of the interaction effect influence the bias. However, they differ in how the center size and the variance of the center-specific L-distribution affect the bias.

Interestingly, for the directly standardized risk the bias will be larger for relatively small centers, while for the indirectly standardized risk the bias will be larger for relatively large centers. The different impact of center size for direct and indirect standardization is due to the different extrapolation. The largest center contributes most in estimating the working model parameters, resulting in a smaller bias on the regression line for the largest center (Figure 3.2, middle panel). Indeed, in Figure 3.1 the working model does not fit well for the smallest center especially for large values of L (e.g. age) resulting in large bias for the directly standardized risk for that center. For the large center on the other hand, we see little bias on the directly standardized risk. In contrast, for indirect standardization in Figure 3.1 the smallest center extrapolates to a region where we have good fit, resulting in small bias, while it is the other way around for the large center. For the small center, the expected risk under its own care level is approximately correct anyway and the risk for these patients under the care level of the other center is only slightly biased. So the expected risk (3.7) for this center also has small bias as we average the expected risks for that center's patients over all observed care levels.

The impact of the variance of the center-specific L-distribution is best understood when looking at two centers (3.11, 3.12). First, the larger the variance in a center's patient mix is, the greater this center's influence on the estimates of the working model parameters. Second, when for a given center the patient mix has large variance, the performance of the other center will be extrapolated to more extreme values of L for which there may be no good fit (Figure 3.2, left or right panel). Then, due to different extrapolation, the bias on the directly standardized risk for a given center increases with smaller variance in its patient distribution

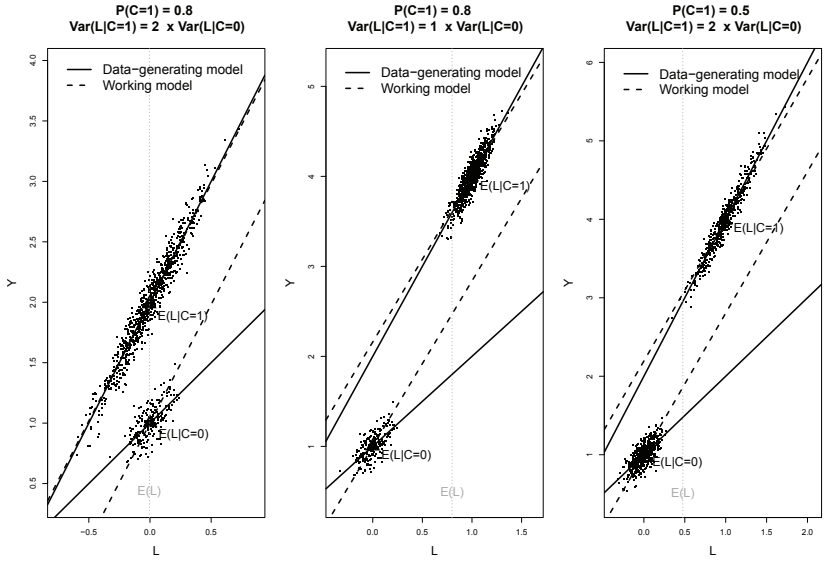


Figure 3.2: Regression line for the data-generating model (3.1) and for the working model (3.8) considering 2 centers (center 0 at the bottom and center 1 on top) and a scalar L.

compared to other centers, while for indirect standardization the opposite is true.

In general, the bias on the indirectly standardized risk will be smaller than on the directly standardized risk and this difference will be more apparent when there are many centers. This because for indirect standardization we take an average over all centers in (3.10), cancelling out positive and negative values, but not for direct standardization in (3.9).

In summary, when ignoring interactions between center and a patient covariate L we only expect large bias in the presence of large interaction effects and large differences in the center-specific mean of L across centers for both standardization techniques. Moreover, for direct standardization the bias is expected to be the largest for the smallest centers and centers with small variance of the center-specific L-distribution. For indirect standardization, the bias is expected to be the largest for the largest centers and centers with large variance of the

center-specific L-distribution.

In the Appendix (Section 3.A) we investigate whether bias can be reduced by using model-based estimators for $E(Y)$ and $E\{Y(c)|C=c\}$. We find that in comparisons with the directly standardized risk $E\{Y(c)\}$ it is not always beneficial to use the model-based overall mortality. Bias on the indirectly standardized risk will never be reduced by using the model-based estimator for $E\{Y(c)|C=c\}$.

3.4 Simulation Study

We perform a simulation study to assess the impact of ignoring interactions in logistic regression models. Besides studying the bias, we also examine efficiency in terms of the mean squared error on the standardized risk of interest.

We simulate $S = 500$ datasets with $n = 10\,000$ patients distributed over $m = 50$ centers and including a scalar patient-specific covariate L. We first generate the patient characteristic, e.g. scaled age, following a standard normal $N(0,1)$ or right-skewed Beta$(1,6)$ distribution and assign each patient to a specific center c, following the propensity score model

$$P(C=c|L) = \frac{\exp(\alpha_{0c} + \alpha_{1c}L)}{\sum_{j=1}^{m}\exp(\alpha_{0j} + \alpha_{1j}L)}, \tag{3.13}$$

where α_{0c} determines the relative center size. Differences in patient mix are large when we impose strongly varying values of α_{1c} among centers. For this study population we generate a binary outcome Y, e.g. 30-day mortality, following a logistic outcome regression model as in (3.1) with the logistic link function. In the Appendix (Section 3.C) we plot the mortality risk in function of L for each center to illustrate the magnitude of the interaction effect. There we also describe the center-specific distribution of the marginally normal standardized or beta distributed covariate L with small or large differences across centers for one simulated dataset.

Bias and precision for the standardized risks are estimated for a working model which includes interactions between L and C or not. Model parameters are estimated using Firth corrected maximum likelihood methods. For conve-

nience, we denote the directly or indirectly standardized risk for center c based on the data-generating model by f_c, based on the fitted working model in simulation run $s = 1, \ldots, S$ by $\hat{f}_c^s$ and the average of the latter over all simulation runs by $\hat{f}_c$. The center-specific bias on the directly or indirectly standardized risk for center c is then estimated by

$$S^{-1} \sum_{s=1}^{S} \left(\hat{f}_c^s - f_c \right) = \hat{f}_c - f_c. \tag{3.14}$$

To prevent that positive bias in some centers is cancelled out by negative bias in other centers, we square these center-specific biases before taking the average over all centers. Then, the overall bias is defined as:

$$\sqrt{m^{-1} \sum_{c=1}^{m} \left(\hat{f}_c - f_c \right)^2 - S^{-1} \hat{\mathrm{Var}} \left(\hat{f}_c^s \right)}, \tag{3.15}$$

which includes a penalty because the average of the estimated squared bias is partly influenced by the imprecision of the estimates. Here, $\hat{\mathrm{Var}}(\hat{f}_c^s)$ denotes the estimated variance of the center-specific standardized risk over all simulation runs. The square root of the overall mean squared error is estimated by:

$$\sqrt{m^{-1} \sum_{c=1}^{m} S^{-1} \sum_{s=1}^{S} \left(\hat{f}_c^s - f_c \right)^2}. \tag{3.16}$$

Finally, we measure the variability in patient mix across centers by the variance of the random intercepts in a random intercept model for L conditional on center.

For normally distributed L the simulation results are similar to the earlier theoretical findings in section 3.3 (Figure 3.3a). That is, when there is little confounding by L, the models with and without interactions give comparable standardized risks. However, when patient mix differs much across centers, the overall bias and root mean squared error for the directly standardized risk $E\{Y(c)\}$ are large. For indirect standardization the overall bias is not as large as for direct standardization, as explained above, and the MSE seems insensitive to excluding the interactions. Here, the variance has by far the largest contribution in the MSE, so apparently precision on the indirectly standardized risk barely

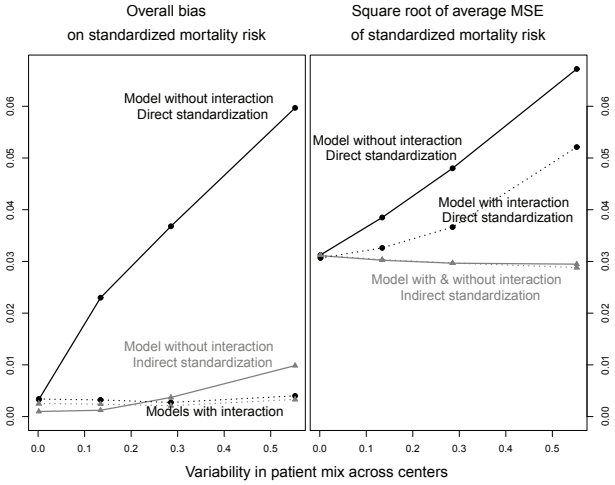
changes when ignoring the center-patient interactions. The overall bias following the model with interactions is sometimes larger than without interactions, which may be due to small sample bias or overfitting problems.

In the Appendix (Section 3.C) we show results for one simulated dataset and indeed detect most bias when there are large differences in patient mix. For directly standardized risks the smallest centers suffer most from bias when ignoring center-patient interactions, while for indirectly standardized risks the largest centers may still suffer substantial bias. It can also be seen that the direction of this bias is not necessarily so that it shrinks the estimated outcome more towards the overall mean or zero, which would mask centers from being detected as having outlying performance. In practice, we thus do not know a priori whether the center's performance is over- or underestimated due to ignoring the interactions.

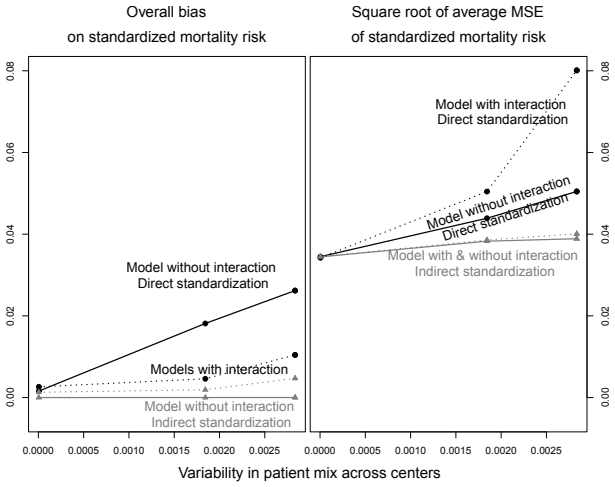
Surprisingly, for beta distributed L in Figure 3.3b we see less bias for the model without interactions than with interactions for indirect standardization. It is also remarkable that for direct standardization and large variability in patient mix, the MSE is larger for the model with interactions than for the model without interactions. Both these findings are due to fitting problems when modelling interactions with beta distributed L.

3.5 Data Analysis: Riksstroke, the Swedish National Quality Register for Stroke Care

Riksstroke (http://www.Riksstroke.org/eng) is a national quality register for acute stroke, collecting data from all 90 Swedish hospitals. The register contains 249 414 adult patients ($\geq$ 18 years) with first registered stroke between 2001 and 2012. We consider patients diagnosed with ischemic stroke (ICD-10 I63), intracerebral haemorrhage (ICD-10 I61) or unspecified acute cerebrovascular event (ICD-10 I64). Centers are compared in terms of directly or indirectly standardized 30-day mortality risks that correct for the patients' sex, age, level of consciousness at arrival (alert, drowsy or unconscious) which is a proxy for baseline severity and time to hospital (hours between stroke and arrival at hospital). The latter



(a) Standard normal distribution for L.



(b) Beta distribution for L.

Figure 3.3: Estimated bias and precision for direct and indirect standardization are based on $S = 500$ simulations. Black dots are used for direct standardization and gray triangles for indirect, full lines are used for models without interactions and dotted lines for models with interactions.

could be an important predictor because brain tissue is rapidly lost as stroke progresses and the sooner treatment (e.g. thrombolysis) is initiated, the larger the probability of a favorable outcome (The ATLANTIS, ECASS, and NINDS rt-PA Study Group Investigators, 2004). The observed time to hospital was more than 24 hours for several patients which are thought to be mistakes in the registration and therefore truncated at 24 hours. Interactions with patient's age may arise when some centers make special efforts for the revalidation of older patients. Differences in center performance may also differ across groups of time until hospital arrival depending on differences in prenotification systems (Lin et al., 2012) and time from arrival to thrombolysis treatment. Riksstroke typically reports directly standardized risks as one aims to compare intrinsic qualities of the centers. Here, we will estimate directly and indirectly standardized risks as we aim to provide insight in the bias in the setting studied here, where each of both standardization techniques could be of interest, depending on the research question posed.

There are 2 852 records with missing consciousness level and 148 910 with missing time to hospital. We discuss the results of two different approaches to handle these missing data: (1) We assume that the data are missing completely at random (MCAR) and perform a complete case analysis. To prevent quasi-complete separation, we also exclude the 2 smallest centers (center size 5 and 22) with respectively 0 and 1 death within 30 days after admission. This resulted in a reduction of the dataset to 100 207 records and overall 30-day mortality risk decreased from 13.13% to 12.48%. (2) We assume that the data are missing at random (MAR) and perform 5 imputations of the missing data using the R-package MICE (Buuren and Groothuis-Oudshoorn, 2011). A description of the predictors that were used for the imputation models is given in the Appendix (Section 3.C). As we need to allow for interactions with center in the outcome model, we fit separate imputation models per center, with center sizes ranging from 56 to 11 669 (Median 2 324). No outcome values were missing.

Standardized outcomes were based on a Firth corrected fixed effects model, with or without interactions between center and time to hospital or age. To suggest a functional form for time to hospital in the outcome regression model while accounting for the other prognostic factors, we categorized time following

its 10% quantiles in the model without interactions. We found a good fit for a loglinear effect of time, and similarly for a linear and quadratic age effect. We found that for longer time to hospital the 30-day mortality risk decreased for alert patients, while for drowsy or unconscious patients the risk increased (Figure 3.14 in Appendix). Therefore we will allow for an interaction between time to hospital and consciousness level in all fitted outcome models. The decreased risk for alert patients may be due to different baseline severity within this group of patients: Patients with a less severe stroke have lower mortality risk but it also takes longer to recognize the symptoms and reach the hospital, while patients with obvious symptoms arrive earlier but also have a higher mortality risk. For drowsy or unconscious patients the symptoms are more apparent, and patients who arrive early will have a lower mortality risk than those who arrive late.

In summary we fit three models, the first includes the center where the patient was treated, patients' sex, age and quadratic age, level of consciousness and log transformed time to hospital as main effects and an interaction between consciousness and time to hospital. The other two models additionally allow the center effect to depend on the linear effect of age or the loglinear effect of time to hospital.

An overall Wald test for interactions with center was obtained for age (p-value 0.009 for CC and < 0.001 for MI) and for time to hospital (p-value < 0.001 for CC and MI). We will now investigate how the standardized risks differ when based on a model with or without interactions between center and patient's age or time to hospital. Results are based on the complete cases (CC) or the multiple imputed data (MI), where for the latter we report results averaged over 5 imputed datasets unless otherwise stated.

We see substantial differences in patient mix across centers coming from time to hospital (Figure 3.4 for MI, Figure 3.12 for CC in Appendix) and only minor differences are seen for age. We measure the variability in patient mix across centers as before, i.e. by the variance on the random intercepts in a random intercept model for L conditional on center, and averaged over the imputed datasets we obtain 0.022 (CC) or 0.024 (MI) for standardized log time to hospital compared to 0.017 (CC) or 0.013 (MI) for standardized age. So from previous theoretical findings we know that the model ignoring interactions with time

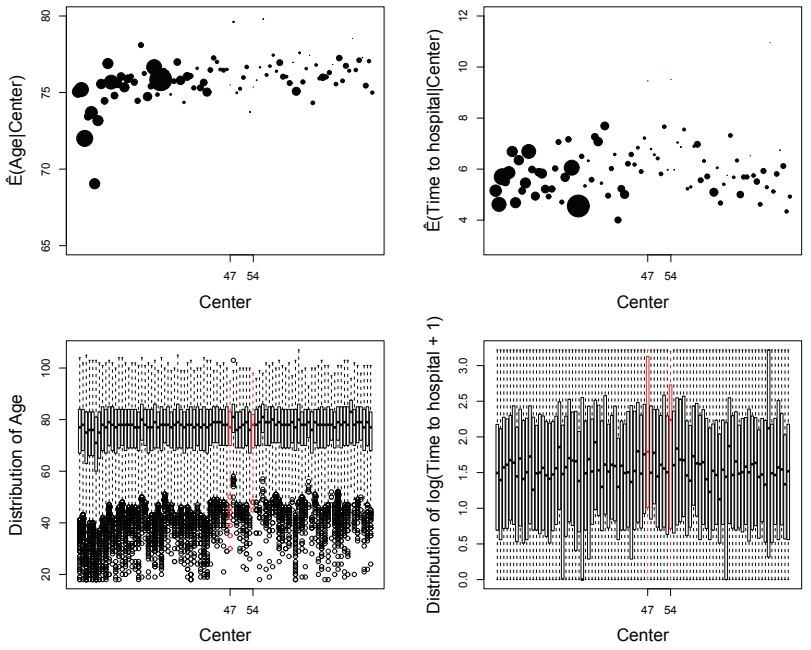


Figure 3.4: Center-specific values for the patient's age and time to hospital (hours) ***for one imputed dataset.*** *Bubble size is proportional to center size. Center 47 and 54 have more than* 1% *difference in its estimated potential full population risk when ignoring interactions with time to hospital (MI).*

to hospital rather than age may induce larger bias on the standardized risks, although the bias will be minor for both (Figure 3.3a).

In general, we see negligible differences in standardized risk when based on a model without or with interactions between center and either age or time to hospital (Figure 3.5 for MI, Figure 3.13 for CC in Appendix). However, for the direct standardization we found 2 (CC) or 2 (MI) centers with a difference of more than 1% in risk when ignoring interactions with time to hospital and 1 (CC) center when ignoring interactions with age. As expected, these differences are larger for direct compared to indirect standardization (Table 3.1). In addition, these differences are larger for time to hospital than for age. So, although for

	Max. difference (%)		Average difference (%)		No. centers with difference > 1%	
	CC	MI	CC	MI	CC	MI
Direct standardization						
Age	1.31	0.63	0.26	0.14	1	0
Time to hospital	1.33	2.83	0.28	0.35	2	2
Indirect standardization						
Age	0.22	0.24	0.05	0.04	0	0
Time to hospital	0.37	0.12	0.09	0.03	0	0

Table 3.1: The difference in estimated standardized risk between the model with and without interactions between center and patient's age or time to hospital, based on complete cases (CC) or multiple imputed data (MI). We report the maximum difference, the average difference (square root of the average of squared differences) and the number of centers for which the difference in standardized risk exceeds 1%.

some centers we found a strong interaction with age, the standardized risks were found to be more robust because the age distribution does not differ much across centers.

3.6 Discussion

We found that if some centers actually perform better on a specific group of patients compared to other centers, then ignoring this in the analysis may bias the directly and indirectly standardized risks when the corresponding patient characteristic is very differently distributed between centers, but bias is negligible otherwise. We therefore advise special attention to interactions with covariates whose distribution differs substantially across centers. When there is no large variability in patient mix, then the common practice of ignoring center-patient interactions does not severely impact standardized mortality risks. In general we notice larger bias for directly standardized compared to indirectly standardized risks. However, for directly standardized risks the largest bias is seen for centers with the smallest proportion of registered patients as opposed to the larger centers for indirectly standardized risks. In our study we found the same trends for the overall root mean squared error. Of course the interaction effect will need

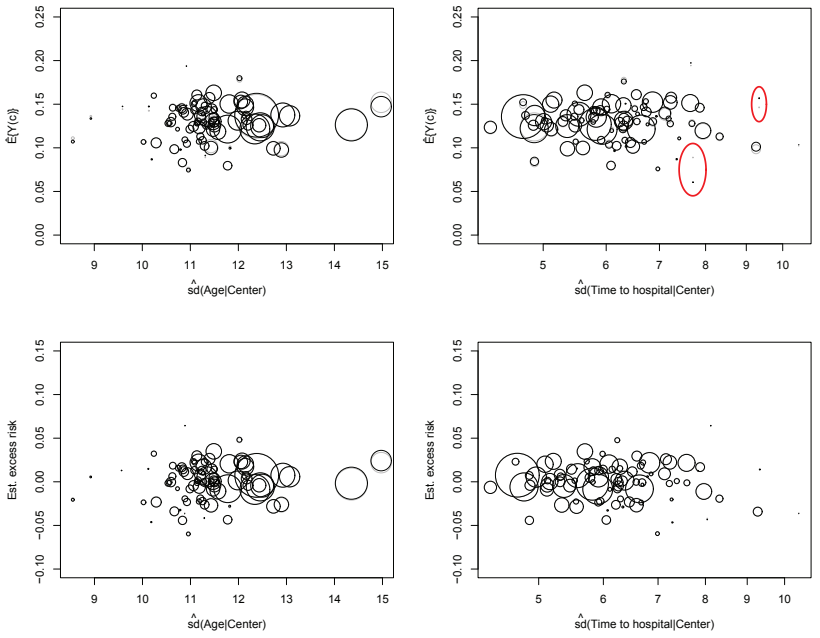


Figure 3.5: The directly or indirectly standardized risk per center, with or without interactions between center and patient's age or time to hospital (grey without and black with interactions), in function of the standard deviation of the center-specific distribution of patient's age or time to hospital ***for multiple imputation analysis****. Bubble size is proportional to center size and ellipses indicate centers with more than* 1% *difference in estimated mortality risk.*

to claim its role when interest lies in prediction of the mortality risk for a specific subgroup rather than directly or indirectly standardized risks.

To detect centers with low or high mortality risks, we have applied a similar center classification technique as in (Varewyck et al., 2014) on the simulated data (See Appendix Section 3.B). A center is classified as low/high risk if the data provide sufficient evidence that the standardized risk exceeds a clinical benchmark e.g. relative to the population average risk $E(Y)$. We found that ignoring center-patient interactions has similar impact on correct center classification as on the bias: the largest differences are seen for direct standardization and large differences in case mix across centers (See Appendix Section 3.C). Surprisingly, the power to detect outlying performance is not always decreased by mistakenly ignoring the interactions, but then we see more centers wrongly classified as having outlying performance. Reassuringly, in general the percentage of correct center classification is very similar for the model with and without interactions and this both for direct and indirect standardization.

We expect that the impact of center-patient interactions on the standardized risk depends on the considered disease. For example, for a register on a non-acute surgical procedure we may expect a large impact. First, patient mix may differ substantially across hospitals when patients can choose the hospital where they are treated. Moreover, treatment and thus the patient's mortality risk is partly based on the surgeons' decisions and experience, which makes it more subject to effect modification, e.g. when some hospitals are less experienced with a specific surgery. On the other hand, for a register on acute stroke, patients are mostly treated at the nearest hospital so there is less confounding. Furthermore, well-defined treatment guidelines for this disease result in the same procedure given in each clinical center, thus we expect the difference in mortality risk between e.g. old and young patients to be similar across centers.

In practice it is not always possible to estimate all interaction parameters, especially when the number of patient characteristics is large. One option is to use prior knowledge on hospital specialization and reduce the factors for which interactions may be considered. In addition or alternatively a summary measure for the patient's baseline severity may reduce the number of parameters to be estimated, e.g. in the form of propensity scores or prognostic scores (Rubin, 1997;

Hansen, 2008). We used penalized likelihood estimation, more specifically the Firth correction to overcome fitting problems. In this context random effects models are often used which help to reduce the effective model dimension when allowing for differential effects of patient characteristics across centers (Normand et al., 1997; Austin et al., 2003). However, it has been repeatedly shown that the power for detecting outlying center performance is much lower when using normal random center effects compared to Firth corrected fixed effects (Varewyck et al., 2014; Kalbfleisch and Wolfe, 2013). In future work it may be of interest to investigate whether more general regularization methods bring a solution.

3.A Asymptotic Bias Calculation

Following a similar principle as in Liu and Gustafson (2008), we calculate the bias when ignoring interactions between center and one patient characteristic L in a linear regression model, respectively with (main text, 3.1) and without (main text, 3.8) interaction between center and L.

We denote the vector of main effects by $\mathbf{X} = (I(C=1),\ldots,I(C=m))'$ and the vector of interactions by $\mathbf{W} = (L\,I(C=1),\ldots,L\,I(C=m))'$. Let $\mathbf{T} = (\mathbf{X}',\mathbf{W}')'$ be the random vector of covariates in the data-generating model (main text, 3.1), let $(\psi',\beta') = (\psi_1,\ldots,\psi_m,\beta_1,\ldots,\beta_m)$ be the model parameters in the data-generating model and let $\mathbf{S} = (L,\mathbf{X}')'$ be the covariates in the working model (main text, 3.8) ignoring the interaction between center and patient characteristic L.

3.A.1 Direct standardization

For the directly standardized risk, we also define for a given center c the vector $E(\tilde{\mathbf{S}}_c) = (E(L),0,\ldots,1,\ldots,0)'$ of length $1+m$, where the number 1 occurs at position $1+c$. Based on the data-generating model, the potential full population risk for center c is given by:

$$E\{Y(c)\} = \psi_c + \beta_c E(L).$$

Let

$$\begin{aligned} r &= E(L^2) - E(L\mathbf{X}')E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X}) = \sum_{j=1}^{m} P(C=j)\mathrm{Var}(L|C=j) = E\{\mathrm{Var}(L|C)\} \\ \mathbf{V}' &= E(L^2\mathbf{X}') - E(L\mathbf{X}')E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X}\mathbf{X}') \\ &= \Big(P(C=1)\mathrm{Var}(L|C=1)\ \ldots\ P(C=m)\mathrm{Var}(L|C=m)\Big) \end{aligned}$$

Ignoring the interactions, the estimator for $E\{Y(c)\}$ has expected value:

$$\begin{aligned} E_r\{Y(c)\} &= E(\tilde{\mathbf{S}}_c')\{E(\mathbf{S}\mathbf{S}')\}^{-1}E(\mathbf{S}\mathbf{T}')\begin{pmatrix}\psi\\ \beta\end{pmatrix} \\ &= (E(L)\,0\ldots1\ldots0)\begin{pmatrix}E(L^2) & E(L\mathbf{X}')\\ E(L\mathbf{X}) & E(\mathbf{X}\mathbf{X}')\end{pmatrix}^{-1}\begin{pmatrix}E(L\mathbf{X}') & E(L\mathbf{W}')\\ E(\mathbf{X}\mathbf{X}') & E(\mathbf{X}\mathbf{W}')\end{pmatrix}\begin{pmatrix}\psi\\ \beta\end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
&= (E(L)\,0\ldots1\ldots0) \\
&\quad \times \begin{pmatrix} \frac{1}{r} & -\frac{1}{r}E(L\mathbf{X}')E(\mathbf{X}\mathbf{X}')^{-1} \\ -\frac{1}{r}E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X}) & E(\mathbf{X}\mathbf{X}')^{-1} + \frac{1}{r}E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X})E(L\mathbf{X}')E(\mathbf{X}\mathbf{X}')^{-1} \end{pmatrix} \\
&\quad \times \begin{pmatrix} E(L\mathbf{X}') & E(L^2\mathbf{X}') \\ E(\mathbf{X}\mathbf{X}') & E(L\mathbf{X}\mathbf{X}') \end{pmatrix} \begin{pmatrix} \psi \\ \boldsymbol{\beta} \end{pmatrix} \\
&= (E(L)\,0\ldots1\ldots0) \begin{pmatrix} \mathbf{0}^T & \frac{1}{r}\mathbf{V}' \\ \mathbf{I}_m & -\frac{1}{r}E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X})\mathbf{V}' + E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X}\mathbf{X}') \end{pmatrix} \begin{pmatrix} \psi \\ \boldsymbol{\beta} \end{pmatrix} \\
&= E(L)\frac{1}{r}\mathbf{V}'\boldsymbol{\beta} + \psi_c \\
&\quad + (0\ldots1\ldots0)\left\{-\frac{1}{r}E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X})\mathbf{V}' + E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X}\mathbf{X}')\right\}\boldsymbol{\beta} \\
&= E(L)\frac{1}{r}\mathbf{V}'\boldsymbol{\beta} + \psi_c + (0\ldots1\ldots0)\left\{-\frac{1}{r}\begin{pmatrix} E(L|C=1) \\ \vdots \\ E(L|C=m) \end{pmatrix}\mathbf{V}'\right. \\
&\quad \left. + \begin{pmatrix} E(L|C=1) & 0 & \cdots & 0 \\ 0 & E(L|C=2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & E(L|C=m) \end{pmatrix}\right\}\boldsymbol{\beta} \\
&= \{E(L) - E(L|C=c)\}\frac{\mathbf{V}'\boldsymbol{\beta}}{r} + \psi_c + \beta_c E(L|C=c).
\end{aligned}
$$

Then bias on the directly standardized risk for center c due to ignoring the center-patient interaction can be written as:

$$
\begin{aligned}
E\left[E_r\{Y(c)\} - E\{Y(c)\}\right] &= \{E(L|C=c) - E(L)\}\left\{\beta_c - \frac{\mathbf{V}'\boldsymbol{\beta}}{r}\right\} \\
&= \{E(L|C=c) - E(L)\}\left[\beta_c - \sum_{j=1}^{m}\frac{P(C=j)\mathrm{Var}(L|C=j)\beta_j}{E\{\mathrm{Var}(L|C)\}}\right].
\end{aligned}
$$

The asymptotic bias on the directly standardized risk for two centers is illustrated in Figure 3.6.

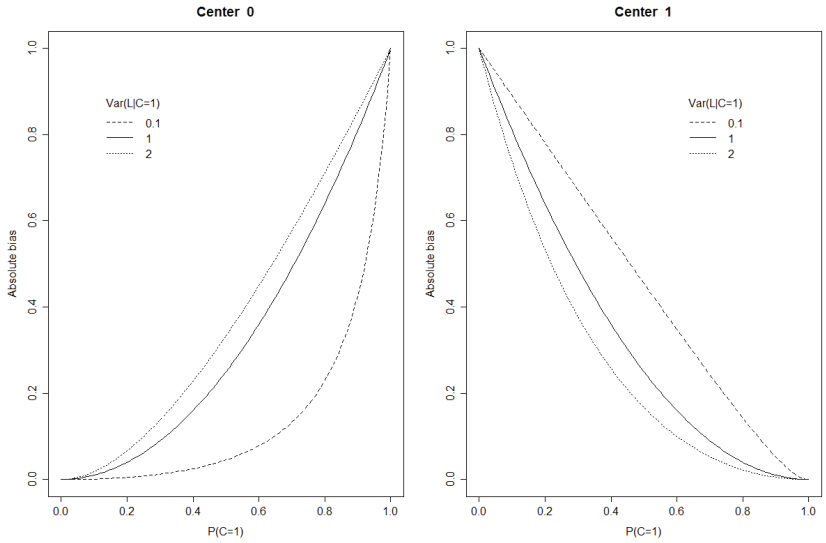


Figure 3.6: Asymptotic bias on the directly standardized risk for 2 centers, when ignoring interaction between center and patient-specific characteristic, following equation (3.11, main text) with $\beta_1 - \beta_0 = 1$, $E(L|C=1) - E(L|C=0) = 1$ *and* $Var(L|C=0) = 1$.

3.A.2 Indirect standardization

For the indirectly standardized risk, we define for a given center c the vector $E(\tilde{\mathbf{S}}_c) = (E(L|C=c), m^{-1}, \ldots, m^{-1})'$ of length $m+1$. Knowing the data-generating model, the center-specific average of expected outcomes in center c is given by:

$$m^{-1}\sum_{c^*=1}^{m} E\{Y(c^*)|C=c\} = m^{-1}\sum_{c^*=1}^{m}\left\{\psi_{c^*} + \beta_{c^*}E(L|C=c)\right\}.$$

However, when ignoring the interactions the fixed effects estimator of the latter has expected value:

$$\begin{aligned}
&m^{-1}\sum_{c^*=1}^{m} E_r\{Y(c^*)|C=c\} \\
&= E(\tilde{\mathbf{S}}_c')\{E(\mathbf{S}\mathbf{S}')\}^{-1}E(\mathbf{S}\mathbf{T}')\begin{pmatrix}\psi\\ \beta\end{pmatrix} \\
&= \left(E(L|C=c)\; m^{-1}\ldots m^{-1}\right)\begin{pmatrix}\mathbf{0}^T & \frac{1}{r}\mathbf{V}' \\ \mathbf{I}_m & -\frac{1}{r}E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X})\mathbf{V}' + E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X}\mathbf{X}')\end{pmatrix}\begin{pmatrix}\psi\\ \beta\end{pmatrix} \\
&= E(L|C=c)\frac{1}{r}\mathbf{V}'\beta + m^{-1}\sum_{c^*=1}^{m}\psi_{c^*} \\
&+ \left(m^{-1}\ldots m^{-1}\right)\left\{-\frac{1}{r}E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X})\mathbf{V}' + E(\mathbf{X}\mathbf{X}')^{-1}E(L\mathbf{X}\mathbf{X}')\right\}\beta \\
&= E(L|C=c)\frac{1}{r}\mathbf{V}'\beta + m^{-1}\sum_{c^*=1}^{m}\psi_{c^*} + \left(m^{-1}\ldots m^{-1}\right)\left\{-\frac{1}{r}\begin{pmatrix}E(L|C=1)\\ \vdots \\ E(L|C=m)\end{pmatrix}\mathbf{V}'\right. \\
&\left. + \begin{pmatrix}E(L|C=1) & 0 & \cdots & 0 \\ 0 & E(L|C=2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & E(L|C=m)\end{pmatrix}\right\}\beta \\
&= \left\{E(L|C=c) - m^{-1}\sum_{c^*=1}^{m}E(L|C=c^*)\right\}\frac{\mathbf{V}'\beta}{r} + m^{-1}\sum_{c^*=1}^{m}\left\{\psi_{c^*} + \beta_{c^*}E(L|C=c^*)\right\}.
\end{aligned}$$

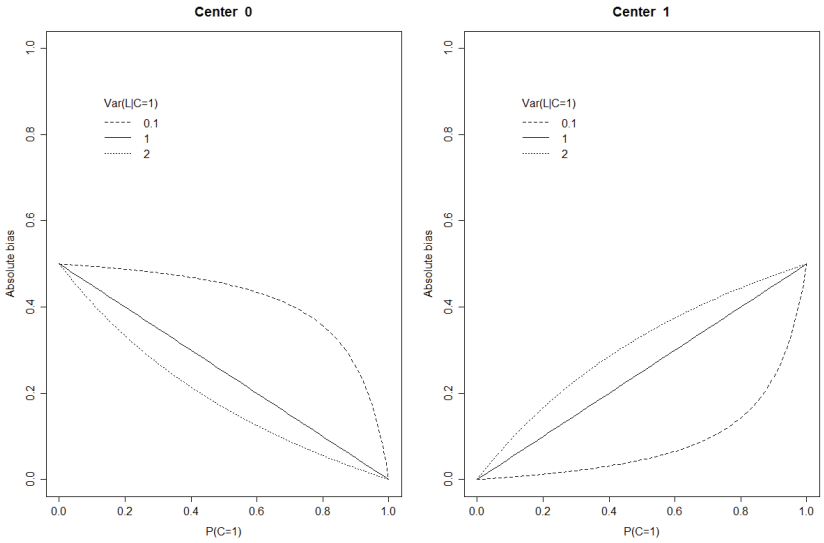


Figure 3.7: Asymptotic bias on the expected outcome in center c (indirect standardization) for 2 centers, when ignoring interaction between center and patient-specific characteristic, following equation (3.12, main text) with $\beta_1 - \beta_0 = 1$, $E(L|C=1) - E(L|C=0) = 1$ *and* $Var(L|C=0) = 1$.

So, bias on the indirectly standardized risk for center c is

$$
\begin{aligned}
& E\left[m^{-1} \sum_{c^*=1}^{m} E_r\{Y(c^*)|C=c\} - m^{-1} \sum_{c^*=1}^{m} E\{Y(c^*)|C=c\} \right] \\
& = m^{-1} \sum_{c^*=1}^{m} \left\{ E(L|C=c^*) - E(L|C=c) \right\} \left[\beta_{c^*} - \frac{\mathbf{V}'\boldsymbol{\beta}}{r} \right] \\
& = m^{-1} \sum_{c^*=1}^{m} \left\{ E(L|C=c^*) - E(L|C=c) \right\} \left[\beta_{c^*} - \sum_{j=1}^{m} \frac{P(C=j)\mathrm{Var}(L|C=j)\beta_j}{E\{\mathrm{Var}(L|C)\}} \right].
\end{aligned}
$$

The asymptotic bias on the indirectly standardized risk for two centers is illustrated in Figure 3.7.

3.A.3 Model-based estimators when comparing risks

In practice, the directly standardized risk $E\{Y(c)\}$ is often compared to the overall mortality. We investigate whether bias is reduced when estimating the latter by the average of the directly standardized risks over all centers $\hat{E}_c[\hat{E}\{Y(c)\}]$ instead of $\hat{E}(Y)$. The bias on $E_c[E\{Y(c)\}]$ due to ignoring center-patient interactions can be calculated as the average of the bias on $E\{Y(c)\}$ over all centers. For two centers, the bias when comparing $E\{Y(0)\}$ with this model-based overall mortality is then

$$\text{Bias}\left(E\{Y(0)\} - E_c[E\{Y(c)\}]\right) = \frac{1}{2}\text{Bias}\left(E\{Y(0)\}\right) - \frac{1}{2}\text{Bias}\left(E\{Y(1)\}\right), \tag{3.17}$$

while $E(Y)$ is unbiased so,

$$\text{Bias}\left(E\{Y(0)\} - E(Y)\right) = \text{Bias}\left(E\{Y(0)\}\right). \tag{3.18}$$

It is not beneficial to use the model-based overall mortality when the absolute value of (3.17) is larger than the absolute value of (3.18). This is the case when

$$\text{Bias}(E\{Y(1)\}) > 3\ |\text{Bias}(E\{Y(0)\})|$$

or, equivalently

$$\begin{aligned} \text{sign}[\text{Bias}(E\{Y(1)\})] &= -\text{sign}[\text{Bias}(E\{Y(0)\})], \text{ and} \\ |\text{Bias}(E\{Y(1)\})| &> |\text{Bias}(E\{Y(0)\})|. \end{aligned}$$

We give an example for each case in Table 3.2.

For indirect standardization, the model-based average of observed risks for a given center c is estimated by:

$$\frac{\sum_{i=1}^{n} g\left(L_i\hat{\beta} + \hat{\gamma}_c\right) I(C_i = c)}{\sum_{i=1}^{n} I(C_i = c)}. \tag{3.19}$$

Now, one of the score equations for Firth corrected maximum likelihood estima-

Bias on $E\{Y(0)\}$	Bias on $E\{Y(1)\}$	Bias on $E\{Y(0)\} - E_c[E\{Y(c)\}]$
0.10	0.40	−0.15
−0.10	−0.40	0.15
0.10	−0.20	0.15
−0.10	0.20	−0.15

Table 3.2: Toy example for the cases when it is not beneficial to use the model-based overall mortality in comparisons with $E\{Y(0)\}$.

tion is exactly

$$\sum_{i=1}^{n} I(C_i = c)\left\{Y_i - g\left(L_i\hat{\beta} + \hat{\gamma}_c\right)\right\} = 0.$$

Then, it follows that the model-based estimator (3.19) for center c is identical to the average of observed risks (main text, 3.6). So, bias on the indirectly standardized risk will not be reduced by using this model-based estimator for $E\{Y(c)|C = c\}$.

3.B Decision Criterion for Labelling Centers

For directly standardized risks, a center is classified as low/high risk if the data provide sufficient evidence that the potential risk $E\{Y(c)\}$ exceeds a benchmark relative to the population average risk $E(Y)$, it is classified as accepted otherwise. So a center is classified as having low risk if

$$\hat{E}\{Y(c)\} + z_k \times \mathrm{sd}\left(\hat{E}\{Y(c)\}\right) < (1 - \lambda_1)\, E(Y)$$

or as high risk if

$$(1 + \lambda_1)\, E(Y) < \hat{E}\{Y(c)\} - z_k \times \mathrm{sd}\left(\hat{E}\{Y(c)\}\right).$$

Here, λ_1 expresses a clinically meaningful tolerance level (e.g. 20%) indicating how much the center-specific potential risk is allowed to depart from the current population average risk $E(Y)$. In practice such envisaged reference is likely to steer the choice of λ_1 once $E(Y)$ is known or has been estimated. Further, z_k is the $k \times 100$th percentile of the standard normal distribution, so k (e.g. 0.75)

expresses the degree of statistical evidence required before flagging a center as low/high risk. In Varewyck et al. (2014) it is explained how $\text{sd}(\hat{E}\{Y(c)\})$ can be estimated.

Similarly for indirectly standardized risks, we will classify a center as having low risk if

$$\hat{\text{Excess}}\text{ risk} + z_k \times \text{sd}\left(\hat{\text{Excess}}\text{ risk}\right) < -\lambda_2$$

or as high risk if

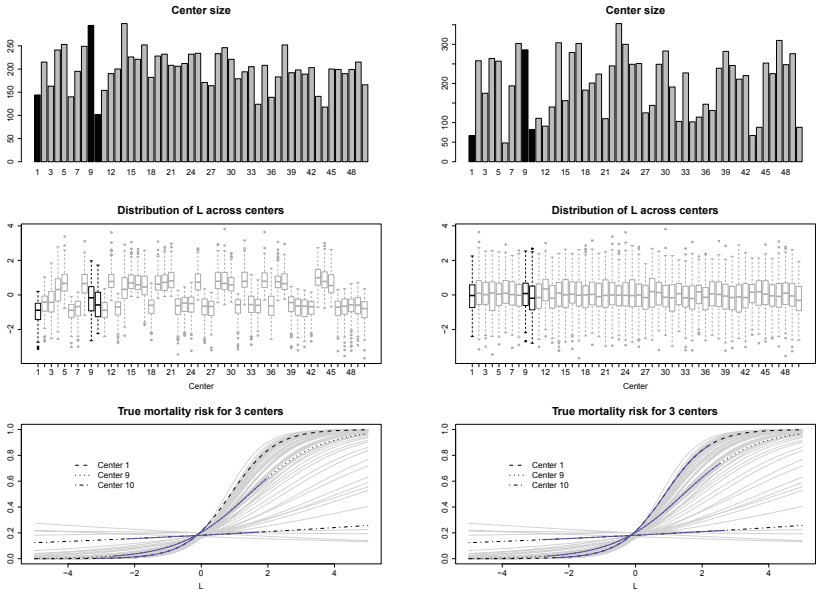
$$\lambda_2 < \hat{\text{Excess}}\text{ risk} - z_k \times \text{sd}\left(\hat{\text{Excess}}\text{ risk}\right).$$

Here again λ_2 expresses a clinically meaningful tolerance level (e.g. 5%) and k (e.g. 0.75) is a measure for statistical evidence. To obtain comparable results between direct and indirect standardization, one can choose $\lambda_2 = \lambda_1 \times E(Y)$ because then the width for acceptance is the same for both standardizations, i.e. $2 \times \lambda_2 = 2 \times \lambda_1 \times E(Y)$.

3.C Additional Results on Simulation Study and Data Analysis

We illustrate the center-specific distribution of the marginally normal standardized or beta distributed covariate L with small or large differences across centers for one simulated dataset in Figure 3.8 and 3.9. As discussed in the paper, we illustrate in Figure 3.10 the mean bias for 1 simulated dataset, for large versus small differences in patient mix and standardized normal L. Results on correct center classification for the simulated data can be found in Table 3.3 for direct standardization and in Table 3.4 for indirect standardization.

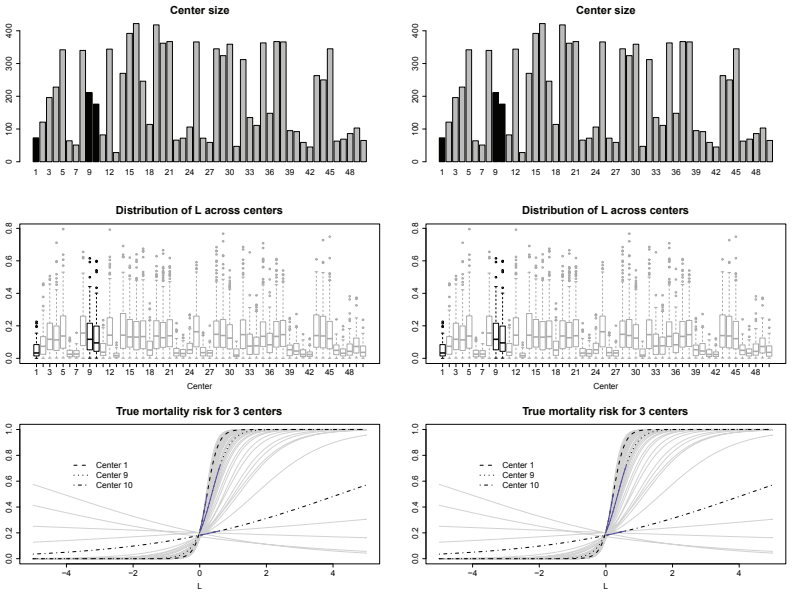
A univariate descriptive analysis of the Riksstroke data can be found in Table 3.5. Additional results for the multiple imputed analysis are shown in Figure 3.11 and for complete case analysis in Figure 3.12 and 3.13. We also studied the effect of time to hospital per consciousness level in Figure 3.14. The estimated interaction effect in the outcome model with interaction between center and time to hospital or age is illustrated in Figure 3.15.



(a) Scenario with large differences in patient mix and standardized normal L.

(b) Scenario with small differences in patient mix and standardized normal L.

Figure 3.8: Illustration of center sizes, variability in patient mix and mortality risk highlighting 3 specific centers in black with full line for their range of L (1 simulated dataset).



(a) Scenario with large differences in patient mix and beta distributed L.

(b) Scenario with small differences in patient mix and beta distributed L.

Figure 3.9: Illustration of center sizes, variability in patient mix and mortality risk highlighting 3 specific centers in black with full line for their range of L (1 simulated dataset).

L-distribution			$N(0,1)$			
Patient mix	Small		Medium		Large	
$L \times C$ interaction	no	yes	no	yes	no	yes
Power (%)						
to detect High	42	44	**33**	**43**	32	34
to detect Low	31	29	33	29	**38**	**28**
Type I error (%)						
Accepted as Low	3.2	2.8	4.2	2.8	**13**	6.3
Accepted as High	5.2	5.3	8.8	6.2	**18**	9.1
Serious type I error (%)						
Low as High	0	0	0	0	1.7	0.3
High as Low	0.2	0.1	0.5	0.1	**13**	2.6
Center classification (%)						
correct (L-A-H)	81	82	77	81	**63**	**75**
L-distribution			$Beta(1,6)$			
Patient mix	Small		Medium		Large	
$L \times C$ interaction	no	yes	no	yes	no	yes
Power (%)						
to detect High	47	47	40	44	**33**	**42**
to detect Low	59	57	52	54	44	52
Type I error (%)						
Accepted as Low	1.0	0.9	2.3	2.8	3.0	6.7
Accepted as High	6.6	6.9	7.8	8.2	7.5	**13**
Serious type I error (%)						
Low as High	0	0	0.4	0	0.1	0.6
High as Low	0	0	0.3	1.3	1.0	3.2
Center classification (%)						
correct (L-A-H)	74	74	70	71	67	66

Table 3.3: Center classification (Low, Accepted or High risk) for direct standardization with $\lambda_1 = 20\%$ and $k = 0.75$. Results are for an outcome regression model without (no) or with (yes) interaction between center and patient characteristic L. The 'Type I error for Accepted as Low' for example denotes the estimated probability that a center with a 'truely' accepted mortality risk is actually classified as having low risk.

L-distribution			$N(0,1)$			
Patient mix	Small		Medium		Large	
$L \times C$ interaction	no	yes	no	yes	no	yes
Power (%)						
to detect High	29	27	43	40	**70**	**61**
to detect Low	35	37	49	52	**68**	**76**
Type I error (%)						
Accepted as Low	4.2	4.5	3.9	4.2	6.9	5.6
Accepted as High	2.5	2.4	3.2	2.9	0.9	0.6
Serious type I error (%)						
Low as High	0	0	0	0	0.03	0
High as Low	0.25	0.25	0	0	0	0
Center classification (%)						
correct (L-A-H)	81	81	83	83	83	83
L-distribution			$Beta(1,6)$			
Patient mix	Small		Medium		Large	
$L \times C$ interaction	no	yes	no	yes	no	yes
Power (%)						
to detect High	36	35	41	38	**38**	**31**
to detect Low	64	65	60	62	59	64
Type I error (%)						
Accepted as Low	1.9	1.9	2.5	2.8	3.7	4.5
Accepted as High	4.2	4.1	4.3	3.8	2.9	2.5
Serious type I error (%)						
Low as High	0	0	0	0	0	0
High as Low	0	0	0	0.08	0	0
Center classification (%)						
correct (L-A-H)	76	76	78	78	78	78

Table 3.4: Center classification (Low, Accepted or High risk) for indirect standardization with $\lambda_2 = 0.20 \times E(Y) \approx 4.5\%$ for $N(0,1)$, $\lambda_2 = 0.20 \times E(Y) \approx 5.8\%$ for $Beta(1,6)$ and $k = 0.75$. Results are for an outcome regression model without (no) or with (yes) interaction between center and patient characteristic L. The 'Type I error for Accepted as Low' for example denotes the estimated probability that a center with a 'truely' accepted mortality risk is actually classified as having low risk.

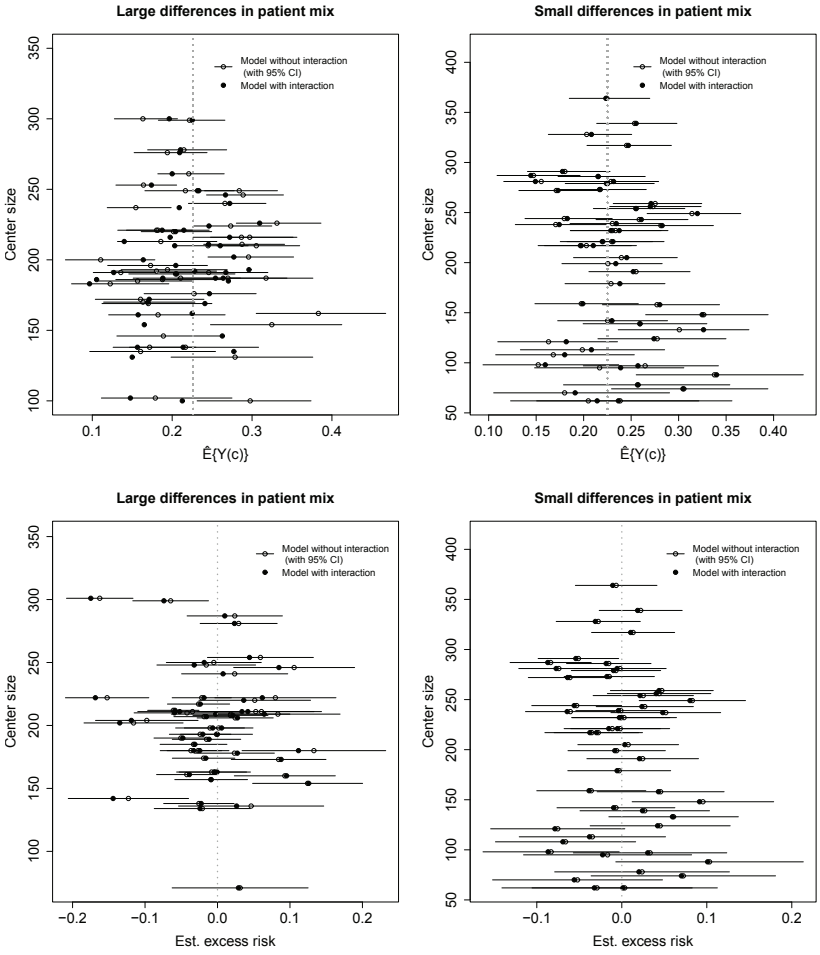


Figure 3.10: Assessing regression to the mean bias for 1 simulated dataset, scenario with large versus small differences in patient mix and standardized normal L. Direct standardization on top, indirect standardization at the bottom.

	Complete cases (n = 83 265)	Missing data (%)	Imputed data (n = 249 414)
Age	76 (67 to 83)	0	78 (68 to 84)
Women	38 694 (46%)	0	123525 (50%)
Consciousness at admission		1	
Alert	70522 (85%)		204621 (82%)
Drowsy	9446 (11%)		31220 (13%)
Unconscious	3297 (4%)		13572 (5%)
Time to hospital (in hours)	2.3 (1.1 to 6.7)	60	3.7 (1.0 to 8.8)
Distance to hospital (in km)	9.6 (2.6 to 23)	1	8.3 (2.5 to 22)
Stroke subtype		0	
Intracerebral haemorrhage (I61)	9676 (12%)		29610 (12%)
Cerebral infarction (I63)	70956 (85%)		210421 (84%)
Unspecified stroke (I64)	2633 (3%)		9383 (4%)
Year of admission		0	
2001	5233 (6%)		19890 (8%)
2002	6186 (7%)		20638 (8%)
2003	5834 (7%)		20915 (8%)
2004	6807 (8%)		21003 (8%)
2005	7970 (10%)		21700(9%)
2006	8056 (10%)		20960 (8%)
2007	8001 (10%)		20446 (8%)
2008	7277 (9%)		20591 (8%)
2009	6795 (8%)		20715 (8%)
2010	7106 (9%)		21225 (9%)
2011	7010 (8%)		20899 (8%)
2012	6990 (8%)		20432 (8%)
Education level		4	
Primary	43391 (52%)		135467 (54%)
Secondary	28079 (34%)		81431 (33%)
University	11795 (14%)		32515 (13%)
Country of birth		1	
Sweden	75200 (90%)		222275 (89%)
Other Nordic	4083 (5%)		12757 (5%)
Other Europe	2835 (3%)		10080 (4%)
Other	1147 (1%)		4301 (2%)
Adjusted yearly income (in 100 SEK)		1	
< 861	8326 (10%)		24864 (10%)
861 to 1334	30497 (37%)		99792 (40%)
2490	34771 (42%)		25058 (40%)
> 2490	9671 (12%)		99700 (10%)
Institutional living	5264 (6%)	1	21769 (9%)
Living alone	34660 (42%)	1	123790 (50%)
Activities of daily living, before admission	6303 (8%)	2	24834 (10%)
Atrial fibrilation	21379 (26%)	2	66661 (27%)
Diabetes	15481 (19%)	1	49658 (20%)
Trt for high blood pressure, before admission	44742 (54%)	2	134526 (54%)
Current smoker	12461 (15%)	13	38422 (15%)
CT-scan after admission	82242 (99%)	0.4	244744 (98%)
Thrombolytic treatment	6323 (8%)	2	8869 (4%)
30-day mortality risk	8571 (10%)	0	32738 (13%)

Table 3.5: Descriptives of the predictors used in the imputation models, based on complete cases or the average over 5 imputed datasets. For continuous variables the median with 1st and 3rd quartile are given and for categorical variables the number of patients (%) are given.

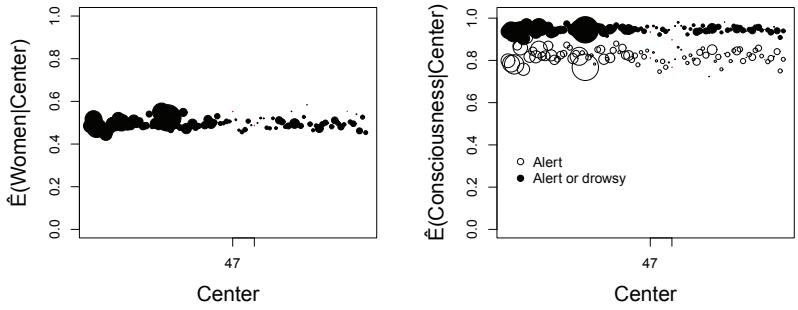


Figure 3.11: Center-specific values for the proportion of female patients and proportion of unconscious or drowsy patients ***for one imputed dataset.*** *Bubble size is proportional to center size. Center 47 and 54 have more than* 1% *difference in its estimated potential full population risk when ignoring interactions with time to hospital (MI).*

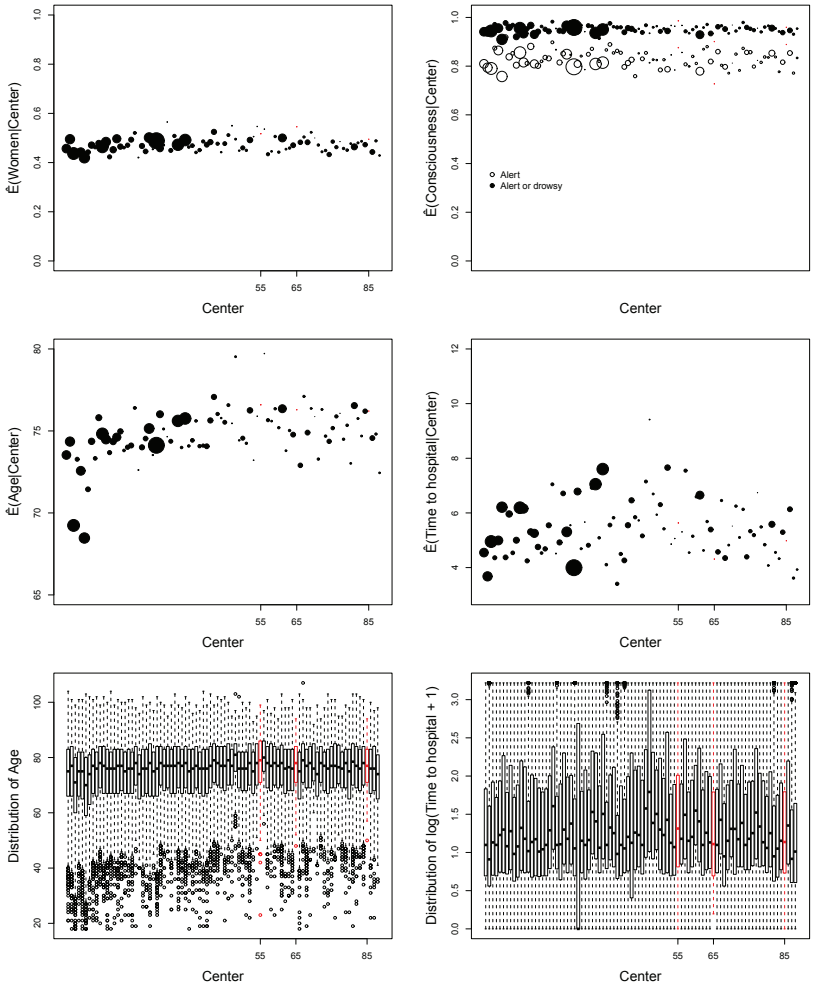


Figure 3.12: Center-specific values for the proportion of female patients, proportion of unconscious or drowsy patients, patient's age and time to hospital (hours) ***for complete cases****. Bubble size is proportional to center size. Center 55 has more than 1% difference in its estimated potential full population risk when ignoring interactions with age, center 65 and 85 for interactions with time.*

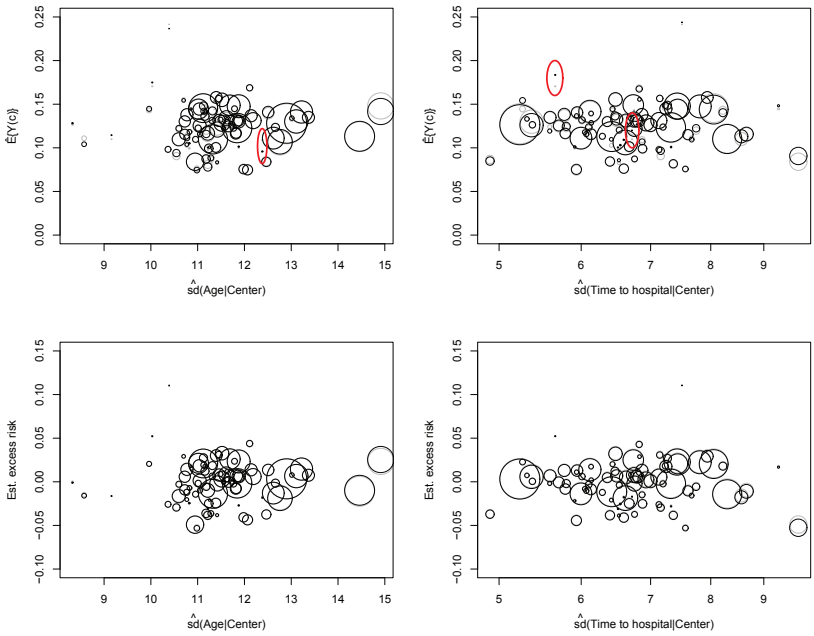


Figure 3.13: The directly or indirectly standardized risk per center, with or without interactions between center and patient's age or time to hospital (grey without and black with interactions), in function of the standard deviation of the center-specific distribution of patient's age or time to hospital ***for complete case analysis****. Bubble size is proportional to center size and ellipses indicate centers with more than* 1% *difference in estimated risk.*

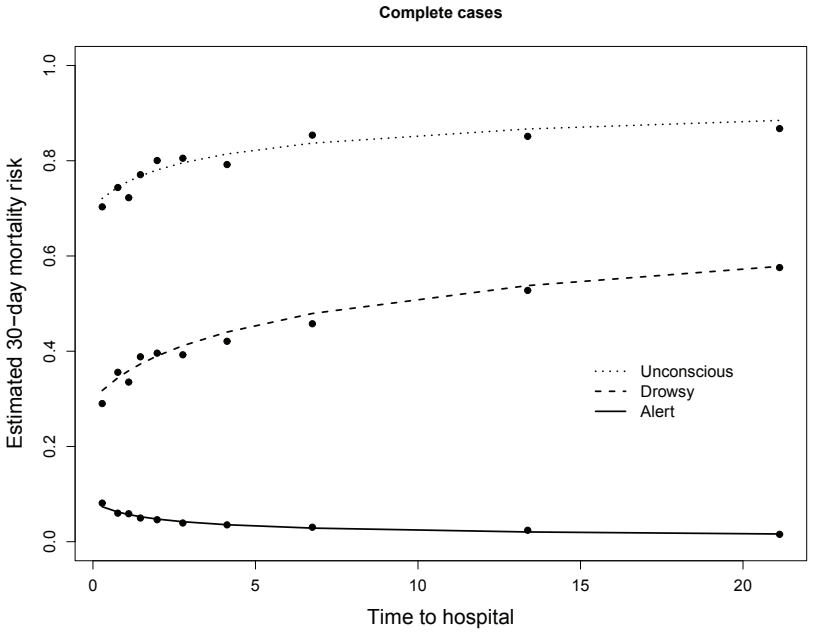


Figure 3.14: The estimated 30-day mortality risk when allowing for a different effect of time to hospital (hours) per consciousness level, considering male patients with mean age and treated at the reference hospital. Time to hospital categories are based on its 10% *percentiles (dots) and fitted lines assume a loglinear effect of time on 30-day mortality risk.*

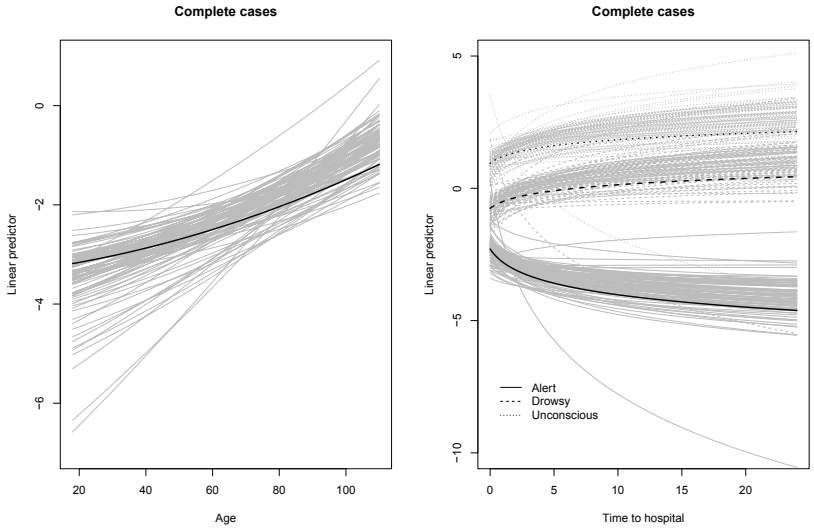


Figure 3.15: Estimated interaction effect in outcome model with interaction between center and time to hospital or age, based on complete cases and for a reference patient (Left: male patient, alert and 0 hours until admission; Right: male patient with mean age).

CHAPTER 4

Cost-efficient Variable Selection for Clinical Registers with Missing Covariate Values

Summary

Clinical registers have expanded enormously and, with that, the hope to capture the important confounders for health-care evaluations and learn more about causal treatment effects. However, more covariates are not always better, because they imply higher measurement costs, possibly less precise and more missing data. We provide a frequentist approach for cost-efficient selection of variables which may be incompletely measured.

We estimate individual mortality risks as well as directly standardized mortality risks based on logistic regression models incorporating patient-specific baseline covariates and added hospital effects. Missing covariate values are handled by complete case analysis or multiple imputation. We search for the subset of patient covariates that suits the budget and minimizes the error on estimated risks. This subset can be approximated by stochastic search algorithms (basic hill-climber or parallel tempering), possibly preceded by the generalized LASSO. For a case study with 83 covariates we found a subset with smaller error than

the Bayesian population RJMCMC approach within a fraction of its search time. For the Swedish register Riksstroke, intrinsically different subsets were found for 30-day mortality prediction versus hospital quality assessment. Analytical and empirical results favored adding consciousness over the more comprehensive but incomplete measure NIHSS of baseline disease severity.

4.1 Introduction

Whether the purpose is personalized medicine or monitoring quality of care, medical registers are rapidly gaining ground. Linkage with existing national registers avoids duplication efforts while maintaining a rich information basis per patient. This in turn enables accurate prediction for (groups of) patients, for example, clinical biomarkers are used to predict which patient populations are more likely to benefit or experience an adverse reaction in response to a given therapy (Trusheim et al., 2007). There are limits however to what can be gained from these additional covariates. The idea that more measured patient characteristics automatically yield better results, not only hits against statistical limitations but also against constraints of measurement costs and human effort involved in registration. First, measurement costs can be lowered by recording fewer characteristics per patient (group). The time and medical infrastructure needed for certain measurements contribute to this cost as well as the potential burden on the patients themselves. Measurement costs should be weighed against the incremental value, on top of predictors that are readily available. Secondly, by reducing the number of predictors one may improve the completeness and accurateness of registration, a recognized weakness of registers (Shahian et al., 2007). The more expensive sickness indicators are thereby expected to be missing more frequently, while basic predictors such as age and gender are accessible for all patients. From a statistical point of view, a full model, including all measured covariates, may suffer overfitting when sample size is limited.

In this paper we search for the subset of covariates that minimizes the error on the predicted individual outcome or on standardized risks for multicenter data, while the total covariate cost is restricted. We estimate these risks based on

logistic regression models that incorporate patient-specific baseline covariates plus hospital effects. Missing data complicate model building, because including patient covariates with missing data may induce selection bias, in view of which the analysis becomes conditional on unverifiable assumptions on the unobserved data. We will consider both complete case analysis and multiple imputation to handle missing data, comparing current practice and statistical preference (Little, 1992). The statistical challenge is even more formidable when evaluating hospital performance, because important confounders of the center-outcome effect should then be selected to avoid confounding bias (Normand et al., 1997; Brookhart et al., 2010). Indeed, if age is for example not accounted for, centers treating mostly older patients will likely show a higher mortality risk, irrespective of their actual care level.

When many covariates are measured, selection is routinely based on automatic forward or backward selection procedures (van der Heijden et al., 2006; Wood et al., 2008), which select variables purely based on statistical significance without taking clinical relevance or measurement costs into account. Instead, we target small prediction errors or accurate performance estimates, depending on the outcome of interest. A characteristic is not only selected for its added value, but also for its low registration cost. We therefore restrict our search to those covariate subsets that respect a predefined (average) budget per patient. Enumerating, not to say evaluating all possible candidate covariate combinations is not manageable even with a modest number of covariates. Therefore, stochastic search algorithms have been suggested in a Bayesian framework (George and McCulloch, 1993), which prove useful in a frequentist context with an adapted definition of a 'better' subset. We define a subset to be better if it yields a smaller cross-validated error on individual predictions or directly standardized risks. We focus on the following two search algorithms: basic stochastic hill-climber and a parallel tempering algorithm (Glover and Kochenberger, 2010).

We apply this variable selection approach to predict 30-day mortality for patients with pneumonia in U.S. hospitals based on data from the RAND corporation (Kahn et al., 1990) and compare results with those reported in Fouskakis et al. (2009). To evaluate hospital performance based on the national register Riksstroke for acute stroke treatment in Swedish hospitals (Asplund et al., 2011),

selection is more complicated as we need to balance the inclusion of confounders needed to avoid confounding bias with the occurrence of missing values which may induce selection bias.

4.2 Methods

4.2.1 Defining the error functions

Let $\mathbf{L}$ be the $(n \times p)$ matrix of p patient-specific baseline characteristics such as gender, age and initial disease status for all n patients. We wish to select a subset $\mathbf{L}_{(\mathbf{S})}$ where $\mathbf{S}$ refers to the vector of inclusion indicators $(S_1, \ldots, S_p)$ such that S_j is 1 if the j-th patient characteristic is included in the outcome regression model and 0 otherwise $(j = 1, \ldots, p)$. Let C be a random variable indicating in which center, out of m centers, the patient was actually treated. In general, we consider the following regression model for outcome Y, e.g. 30-day mortality:

$$E\left(Y|\mathbf{L}_{(\mathbf{S})}, C; \beta_{(\mathbf{S})}, \psi_{(\mathbf{S})}\right) = g\left(\mathbf{L}_{(\mathbf{S})}\,\beta_{(\mathbf{S})} + \sum_{c=1}^{m} \psi_{c,(\mathbf{S})} I(C = c)\right), \tag{4.1}$$

where $\beta_{(\mathbf{S})}$ is an unknown parameter with the dimension of $\mathbf{L}_{(\mathbf{S})}$, $\psi_{(\mathbf{S})} = (\psi_{1,(\mathbf{S})}, \ldots, \psi_{m,(\mathbf{S})})^T$ are the residual center effects and $g(\cdot)$ is a known link function, e.g. the logistic link. The model parameters $(\beta_{(\mathbf{S})}^T, \psi_{(\mathbf{S})}^T)$ will be estimated using the Firth penalized-likelihood method (Firth, 1993), which has been shown to be preferable in this setting (Varewyck et al., 2014), because it avoids undue shrinkage and maintains convergence in the presence of small centers.

We aim to find a subset of patient characteristics, identified by $\mathbf{S}$, that respects the total allowed cost A:

$$\sum_{j=1}^{p} I(S_j = 1)\, a_j \leq A, \tag{4.2}$$

where $a_j \geq 0$ is the cost for measuring and registering covariate j $(j = 1, \ldots, p)$. For the first outcome of interest, the given subset has to minimize the error on the predicted individual outcome, which is defined as:

$$ER_1(\mathbf{S}) = E\left[\left\{E\left(Y|\mathbf{L}^*_{(\mathbf{S})}, C^*; \hat{\beta}_{(\mathbf{S})}, \hat{\psi}_{(\mathbf{S})}\right) - Y^*\right\}^2\right]^{1/2}. \tag{4.3}$$

For a given $\mathbf{S}$, we will first estimate the model parameters $(\boldsymbol{\beta}_{(\mathbf{S})}^T, \boldsymbol{\psi}_{(\mathbf{S})}^T)$ as $(\hat{\boldsymbol{\beta}}_{(\mathbf{S})}^T, \hat{\boldsymbol{\psi}}_{(\mathbf{S})}^T)$ on 80% of the data $(Y, \mathbf{L}_{(\mathbf{S})}, C)$ and then, given the obtained estimates, the prediction error is evaluated on the remaining 20% of the data $(Y^*, \mathbf{L}_{(\mathbf{S})}^*, C^*)$. Using cross-validation techniques, we randomly partition the data $V = 5$ times into two such complementary subsets $(Y, \mathbf{L}_{(\mathbf{S})}, C)$ and $(Y^*, \mathbf{L}_{(\mathbf{S})}^*, C^*)$, which are assumed to have the same distribution. Note that estimating the prediction error using cross-validation may give overly pessimistic results, due to a reduced sample size for parameter estimation (Steyerberg et al., 2001).

In the context of hospital performance, we aim to minimize the error on the estimated directly standardized risks. As the hospital performance measure is intended to represent the causal effect of the given care level, it is important to adjust for confounding (Brookhart et al., 2010). Otherwise, higher mortality risks may be unfairly attributed to a worse care level while they were actually due to differences in patient mix across centers (e.g. initial disease severity status, age). So, let $Y(c)$ indicate the potential outcome for a given patient if treated at the care level of center c. The directly standardized risk, denoted by $E\{Y(c)\}$, encodes the potential full population risk under the care of center c: the risk that would be realized if all patients under study were to experience the care level of that given center c, irrespective of where they were actually treated, e.g. Varewyck et al. (2014). Throughout, we assume that the patient-specific covariates $\mathbf{L}$ are sufficient to adjust for confounding of the center-outcome effect, so that $Y(c) \perp\!\!\!\perp C|\mathbf{L}$ for all c (Hernán and Robins, 2006b). Under this assumption, we have that $E\{Y(c)\} = E\{E(Y|\mathbf{L}, C = c)\}$. Having selected a subset of patient characteristics $\mathbf{S}$, the error at hospital level is defined as:

$$ER_2(\mathbf{S}, c) = E\left[\left(\hat{E}_{(\mathbf{S})}\left\{Y^*(c); \hat{\boldsymbol{\beta}}_{(\mathbf{S})}^*, \hat{\boldsymbol{\psi}}_{(\mathbf{S})}^*\right\} - \hat{E}\left\{Y(c); \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right\}\right)^2\right]^{1/2}, \tag{4.4}$$

where the gold standard $E\{Y(c)\}$ is not known and therefore replaced by the estimated directly standardized risk $\hat{E}\{Y(c); \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\}$ based on half of the data and including all measured covariates. The remaining half of the data is used to estimate the model parameters as $(\hat{\boldsymbol{\beta}}_{(\mathbf{S})}^{*T}, \hat{\boldsymbol{\psi}}_{(\mathbf{S})}^{*T})$ and for a given choice of $\mathbf{S}$:

$$\hat{E}_{(\mathbf{S})}\left\{Y^*(c); \hat{\boldsymbol{\beta}}_{(\mathbf{S})}^*, \hat{\boldsymbol{\psi}}_{(\mathbf{S})}^*\right\} = \hat{E}\left\{E\left(Y|\mathbf{L}_{(\mathbf{S})}^*, C = c; \hat{\boldsymbol{\beta}}_{(\mathbf{S})}^*, \hat{\boldsymbol{\psi}}_{(\mathbf{S})}^*\right)\right\}.$$

This strategy makes sure that estimation of the model parameters and standardization are performed on the same subset of the data, which is common practice for direct standardization. Whilst this approach may result in underestimation of the error $ER_2(\mathbf{S}, c)$ because the gold standard is replaced by the model-based $\hat{E}\{Y(c); \hat{\boldsymbol{\beta}}, \hat{\psi}\}$, it is not expected to change which subset has the smallest error, because the relative ranking of the errors is preserved under the assumption that the estimate of the gold standard is consistent (Brookhart and van der Laan, 2006). We then aim to find the subset $\mathbf{S}$ that meets the cost constraint (4.2) and minimizes:

$$ER_2(\mathbf{S}) = E\{ER_2(\mathbf{S}, c)\}, \tag{4.5}$$

where the average gives equal weight to all centers.

4.2.2 Search methods for cost-efficient variable selection

The complexity of the search strongly differs for the two targeted risks. For individual prediction, we aim to select those covariates that strongly impact individual outcomes, as we target minimization of the prediction error. When interest lies in accurately estimating directly standardized risks, we also need to select the confounders of the center-outcome effect, which is especially important to avoid confounding bias and thus to make fair evaluations of hospital performance. Our methods may fail to select confounders that are strongly correlated with the center choice and only moderately with the outcome. Therefore we suggest to first assess how patient mix differs across centers before the variables are selected. Alternative selection procedures that are better targeted towards the selection of confounders have been suggested in e.g. van der Laan and Gruber (2010); Wilson and Reich (2014); Wang et al. (2012); Vansteelandt et al. (2010), but these do not restrict the total covariate cost. Moreover, in contrast to prediction, risk standardization requires that the regression model is correctly specified, because the fitted model is used to extrapolate results outside the observed covariate range, which is especially the case if patient mix differs strongly across centers. In that case, selecting the subset of confounders that minimizes the error on the directly standardized risks is not sufficient, but one also needs to assume that there is no uncertainty about the structural properties of the outcome regression

model, such as the link function and functional form of the covariates.

In this paper we will consider two stochastic search algorithms to approximate the subset of covariates **S** that minimizes the error functions $ER_1(\mathbf{S})$ and $ER_2(\mathbf{S})$. Notice that we only consider suitable 0/1 combinations for **S**, e.g. including a categorical covariate with dummy coding will change several values of **S** at a time. The basic stochastic hill-climber (Glover and Kochenberger, 2010) starts with a random subset of covariates that meets the covariate cost constraint. In each step of the algorithm, a new neighbor-subset is created by randomly adding, removing or swapping one covariate from the previously accepted subset. It is first checked whether the covariate cost of this new subset exceeds the predefined cost constraint. If so, this subset is not further considered and a new neighbor-subset is created from the last accepted subset. If the cost constraint is met, the outcome model is fitted and the error on the predicted individual outcome or the standardized risk is estimated. If this error is smaller than the currently accepted subset, this new subset is accepted as current subset. If not, the algorithm generates a new neighbor starting from the last accepted subset. This strategy is similar to the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm which allows the number of included covariates to vary during the search (Green, 1995). As the basic hill-climber may be trapped in a local optimum, the procedure is repeated for 10 initial subsets and the best solution over all restarts is chosen. This does not affect the total computation time if these 10 independent restarts are executed on parallel cores of the computer.

Alternatively, we consider the parallel tempering search algorithm, which is superior to the basic stochastic hill-climber as it is able to escape from a local optimum by sometimes accepting subsets whose error is larger than for the current subset solution. For example, when evaluating prediction of the individual's outcome, the probability to accept a new subset solution $\mathbf{S}_{\text{new}}$ over the current solution $\mathbf{S}_{\text{current}}$, for a given temperature t, is:

$$P(\text{accept } \mathbf{S}_{\text{new}}) = \begin{cases} 1 & \text{if } \hat{ER}_1(\mathbf{S}_{\text{new}}) \leq \hat{ER}_1(\mathbf{S}_{\text{current}}) \\ \exp\left\{-\frac{\hat{ER}_1(\mathbf{S}_{\text{new}}) - \hat{ER}_1(\mathbf{S}_{\text{current}})}{t}\right\} & \text{if } \hat{ER}_1(\mathbf{S}_{\text{new}}) > \hat{ER}_1(\mathbf{S}_{\text{current}}), \end{cases} \tag{4.6}$$

and similarly for $ER_2(\mathbf{S})$. The search algorithm runs multiple chains, each with

a given temperature, and exchanges subset solutions between them. Chains with a high temperature thereby allow a lot of freedom to escape from local optima, while those with lower temperatures ease convergence towards an optimum. In our case, the parallel tempering algorithm will concurrently run 5 such Metropolis searches with different temperatures. Solutions are then exchanged to converge to one single best solution. It is advised to perform several Metropolis searches in advance, each with a given temperature, to set the minimum and maximum temperature for the parallel tempering algorithm. In a Bayesian framework, this search strategy has been explored by means of the population-based trans-dimensional RJMCMC algorithm in Fouskakis et al. (2009), which combines ideas from the population-based MCMC and simulated tempering algorithms.

These search algorithms do not terminate internally, so that we have to specify a stopping criterion. Here, we limit the maximum time without finding a better subset. Given the time it takes to evaluate the error for one given subset $\mathbf{S}$, one can calculate the maximum amount of time needed to evaluate all possible neighbors of a given subset solution, which may help to set realistic values for the stopping criterion. The latter may be large if the number of candidate subsets is large or when evaluating the error function for one given subset takes already a considerable amount of time.

In view of computation time, we also investigate the use of the generalized LASSO (Tibshirani and Taylor, 2011), which can serve as a first raw exploration of the variable space. Indeed, the cost constraint in (4.2) suggests the use of a weighted penalty function for LASSO regression. Then, the generalized LASSO estimates are defined by:

$$\underset{(\boldsymbol{\beta},\boldsymbol{\psi})\in\mathbb{R}^{p+m}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \left\{ Y_i - E(Y_i|\mathbf{L}=\mathbf{L}_i, C=C_i;\boldsymbol{\beta},\boldsymbol{\psi}) \right\}^2 + \lambda \left(\sum_{j=1}^{p} a_j|\beta_j| + \sum_{c=1}^{m} b_c|\psi_c| \right) \right], \tag{4.7}$$

where $\lambda \geq 0$ is a tuning parameter and the subscript ($\mathbf{S}=\mathbf{1}$) is omitted as naturally all covariates are considered here. In particular, the center effects have penalty factors $(b_1,\ldots,b_m)$, which can all be set at 1 to penalize the center effects following ordinary LASSO for individual prediction, or close to zero for hospital quality

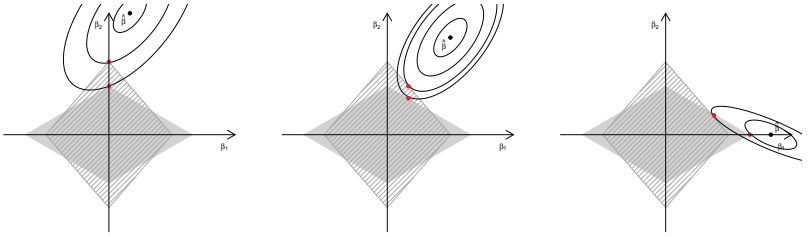


Figure 4.1: The shaded area is the constraint region for equal penalty factors ($|\beta_1| + |\beta_2| \leq 2$, i.e. ordinary LASSO), and the gray colored area for a penalty of β_2 that is twice that of β_1 ($2/3|\beta_1| + 4/3|\beta_2| \leq 2$, generalized LASSO). The ellipses are the contours of least squares error functions which intersect with the constraint regions in the red dots.

assessment so that the center effects are hardly penalized and thus forced into the model. The penalty factors $a_j \geq 0$ for the patient characteristics increase for larger measurement costs. In this way, covariates with a larger cost are intended to be penalized more than cheap ones, reflecting that we are only willing to pay an extra cost if the benefit in terms of smaller errors is substantial. Before estimating the parameters, the penalty factors are internally rescaled to sum to the number of considered model parameters ($p + m$). The generalized LASSO method is also illustrated in Figure 4.1, where compared to the ordinary LASSO, the effect of an expensive covariate is shrunken more and sometimes set to zero while the opposite may be true for the cheaper covariate. We will perform cross-validated (default $V = 10$) generalized LASSO to find the optimal value of λ, i.e. the value that minimizes the mean-squared error on outcome predictions. This value corresponds with a given subset of covariates that not necessarily respects the total cost constraint. Once this subset is known, we will refit the outcome model using Firth corrected maximum likelihood estimation and estimate the cross-validated error, which will allow us to compare the performance of this method with the results from the stochastic search algorithms.

When some covariate values are missing or measured with error, our cost-efficient variable selection approach becomes especially relevant. It allows to evaluate the added value of such a covariate while penalizing it by the cost to fill in missing values or obtain more accurate measurements. Although multiple

imputation has been recommended over complete case analysis (van der Heijden et al., 2006), a combination with variable selection is not straightforward (Wood et al., 2008; Musoro et al., 2014). The completed data sets may result in different best subsets, which are often combined using ad-hoc methods such as retaining those covariates that were selected in at least 50% of the imputations (Vergouwe et al., 2010). In this article, we will obtain one final subset as follows: Given a subset suggested by the search algorithm, we fit the outcome model on each imputed data set. Then, predicted outcomes or standardized risks are averaged across the imputed data sets before the error is evaluated and the suggested subset is retained or rejected. This approach has previously been recommended by Wood et al. (2008), where it is shown that under the 'missing at random assumption' as stated in Little (1992), the obtained Type I error of wrongly including a given variable which has no predictive value, is comparable to what would be achieved if there were no missing data. We have now made this approach computationally feasible, even for large data sets and numerous variables, by using stochastic search algorithms to select the best subset. For the generalized LASSO, we suggest to take the union of the selected subsets over the multiple imputed data.

The search algorithms are run in Java version 1.8.0 using the JAMES framework (De Beukelaer et al., 2015). Statistical analyses are performed in R version 3.1.2.

4.3 Two Case Studies

4.3.1 Subset selection for RAND data

The RAND data contain a representative sample of $n = 2532$ elderly American patients hospitalized in the period 1980-86 with pneumonia, taken from the RAND study (Kahn et al., 1990). An overview of the full set of 83 variables, together with the minutes of abstraction time, ranging from 30 seconds to 10 minutes is given in the Appendix (Table 4.2, taken from Fouskakis et al. (2009)). We will use the basic stochastic hill-climber and the parallel tempering search algorithm to find the subset of covariates that minimizes the error on 30-day mortality

predictions and compare their performance with the population RJMCMC and the generalized LASSO. The data set has no missing values and no information on the center where patients were treated. The estimated average 30-day mortality risk is 15.8%. We perform $V = 5$ cross-validations with 80% + 20% data split, where the selected subjects differ only over the V data splits, but are re-used for each subset selection $\mathbf{S}$ or (restart of the) search algorithm. Calculations are performed on a machine with 3.40 GHz of CPU speed and 8 Gbyte RAM, unless specified otherwise.

The full model, including all 83 covariates has a total cost of 103 and its cross-validated prediction error was estimated to be 0.3162. The computation time to fit this model was only 7.3 seconds. The subset suggested by the RAND committee was based on medical knowledge and includes 14 covariates with total cost 30.5 and $E\hat{R}_1(\mathbf{S}) = 0.3126$. Variable selection for the RAND data has extensively been studied in Fouskakis et al. (2009). For example, when the total cost is allowed to be at most 10, the population RJMCMC method in Fouskakis et al. (2009) selected 8 covariates as best subset with estimated prediction error 0.3179. This search took about 3.3 days, on a machine with 3.66 GHz of CPU speed and 1 Gbyte RAM (Fouskakis et al., 2009). Although differences in prediction error turn out to be negligible, substantial differences in computation time and covariate cost are detected. Clearly, the full model is not preferred as it has a huge covariate cost while its prediction error is not the smallest. Results on the selected subsets are summarized in Table 4.1 and their included covariates are compared in Figure 4.4 (see Appendix).

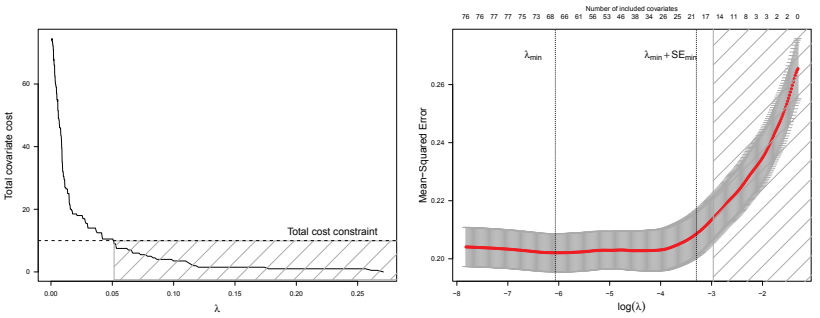
We aim to reduce the computation time and prediction error compared to the population RJMCMC by selecting covariates with different stochastic search algorithms or the generalized LASSO. We constrain the total covariate cost to 10. The maximum time without finding a better subset is set at 10 minutes for the stochastic hill-climber (and descriptive Metropolis searches), while for the parallel tempering algorithm, which is more time-consuming, we double the time to 20 minutes. We performed 100 Metropolis searches to define the minimum (0) and maximum (0.003) temperature for the parallel tempering algorithm (Figure 4.5 in Appendix). The stochastic hill-climber and the parallel tempering search algorithm selected the same subset of 13 covariates as the subset with the

*Table 4.1: Comparison of the variable selection methods on the RAND data. Computation time is for a machine with 3.40 GHz of CPU speed and 8 Gbyte RAM, except for * it was 3.66 GHz of CPU speed and 1 Gbyte RAM.*

Selection method	Prediction error $E\hat{R}_1(\mathbf{S})$	Total cost (constraint)	Computation time	Selected covariates
Full model	0.3162	103 (-)	7.3 secs	83
RAND committee	0.3126	30.5 (-)	0.7 secs	14
Population RJMCMC	0.3179	10 (10)	3.3 days*	8
Basic stochastic hill-climber	0.3039	10 (10)	38 mins	13
Parallel tempering	0.3039	10 (10)	2.2 hrs	13
Generalized LASSO				
with cost constraint	0.3218	9 (10)	9.6 secs	15
at $\lambda_{min} + SE_{min}$	0.3189	13 (-)	9.6 secs	20

smallest estimated prediction error of 0.3039 over 10 restarts. The longest computation time over the 10 independent restarts was 38 minutes for the stochastic hill-climber and 2.2 hours for the parallel tempering algorithm. Our best subset was found only once for the stochastic hill-climber (Figure 4.6 in Appendix) although enough time was given to evaluate all neighbor subset solutions, so that the other restarts were probably trapped in a local optimum. For the parallel tempering algorithm this best subset was found twice, because the algorithm is partially protected against being trapped in a local optimum by concurrently running chains with different temperature. The generalized LASSO with cost constraint selected 15 covariates within a drastically reduced computation time of 9.6 seconds. This subset has an estimated root mean-squared error of 0.4626 and was selected after evaluating 500 values for λ with 10-fold cross-validations. The estimated prediction error for this subset based on Firth corrected fixed effects regression is $E\hat{R}_1(\mathbf{S}) = 0.3218$. Without cost constraint, it can be seen in Figure 4.2 that the subset corresponding with $\lambda = \lambda_{min} + SE_{min}$ has an estimated average mean-squared error (MSE) which exceeds the minimum MSE by only one time its estimated standard error (SE_{min}) and still approximates the cost constraint. This subset solution has a total covariate cost of 13 and $E\hat{R}_1(\mathbf{S}) = 0.3189$, so it is certainly preferred over the one with the smallest mean-squared error, because the latter has a total covariate cost of 67.5.

In summary, the stochastic hill-climber is the preferred method here, be-



(a) Covariate cost *(b) Mean-squared error*

Figure 4.2: Generalized LASSO regression for a range of tuning parameter values λ, for the RAND data. The shaded part meets the total cost constraint. λ_{min} is the tuning parameter corresponding with the minimum mean-squared error and SE_{min} is the estimated standard error of its mean-squared error.

cause it finds the subset with the smallest prediction error that meets the cost constraint, within a relatively short time span. The best subset we found has a slightly smaller error than the subset suggested by the RAND committee and costs only a third. All subsets from the 10 restarts of our stochastic search algorithms have a smaller $E\hat{R}_1(\mathbf{S})$ than the best subset in Fouskakis et al. (2009) (Figure 4.6 in Appendix) and although the reduction in prediction error is at most 1.7%, the reduction in computation time is huge (3.3 days compared to 38 minutes). The smaller prediction error may be explained by the use of a different selection criterion in Fouskakis et al. (2009), namely the Bayesian posterior model probability. Estimating that probability is also very time consuming because 500 000 iterations of the population RJMCMC had to be performed. The generalized LASSO further reduced the computation time, but did not result in smaller prediction errors and did not necessarily meet the total cost constraint. The worse prediction errors may be due to undue shrinkage of the parameter estimates in the fitted model for variable selection while our selection criterion, the cross-validated prediction error, was based on Firth corrected fixed effects regression. If the computation time for the stochastic hill-climber is problematically large, we suggest to first perform a raw exploration of the variable space

with the generalized LASSO. Then in a second phase, the stochastic hill-climber can be applied to further reduce the number of included covariates so that the total cost constraint is met.

4.3.2 Subset selection for Riksstroke

Riksstroke is the Swedish quality register for stroke care which aims to monitor and improve hospital performance and ultimately to ensure the best possible care for stroke patients. We consider 124 308 adult patients (≥ 18 years) with first registered stroke between 2007 and 2012 who were treated in one of 80 Swedish hospitals. The number of registered patients ranges from 103 to 5 832 per hospital. We consider patients diagnosed with ischemic stroke (ICD-10 I63), intracerebral haemorrhage (ICD-10 I61) or unspecified acute cerebrovascular event (ICD-10 I64) and we focus on 30-day mortality as the quality indicator, which is never missing. We focus subsequently on minimizing the error at the patient level $ER_1(\mathbf{S})$ and at the hospital level $ER_2(\mathbf{S})$. In total, 18 baseline patient characteristics are considered, of which some may be incompletely measured (Table 4.3 in Appendix). For example, the covariate NIHSS (National Institutes of Health Stroke Scale) is a comprehensive measure (42 levels) for baseline disease severity but is missing for two-thirds of patients. In contrast, the patient's consciousness level at admission (alert, drowsy or unconscious) is registered for almost all patients.

When some of the included covariates have missing values, the error will be based on either the complete cases (CC) or the multiple imputed (MI) data. For the first method we assume that the missingness indicator $R \perp\!\!\!\perp Y|\mathbf{L}, C$, i.e. the outcome distribution of all data and the complete cases do not differ, given the center and patient characteristics. For the directly standardized risks, we moreover assume that $R \perp\!\!\!\perp \mathbf{L}$, because the average over the distribution of $\mathbf{L}$ is taken and otherwise the intended performance measure would differ for the complete cases and all data. Depending on the included set of covariates, the complete cases left a sample size of at least 41 798 complete records with lower average 30-day mortality risk (8.0% versus 13.1%). To prevent lack of information and quasi-complete separation, we excluded centers with less than 20 registered

patients to fit the model (max. 6 and 9 excluded centers, respectively at patient and hospital level). For the second method, assuming missingness at random we performed 5 imputations of the missing covariate data using the R-package MICE (Buuren and Groothuis-Oudshoorn, 2011). Missing values were imputed using Bayesian linear regression models for continuous covariates, logistic regression for binary covariates and polytomous regression for unordered categorical covariates. Each imputation model includes the 30-day mortality outcome and all baseline covariates, except the one that is imputed. The data are thus multiply imputed before the variable selection procedure is performed. For a given subset **S**, we combine the outcome predictions and the directly standardized risks for each center c ($c = 1,\ldots,m$) over the 5 imputations using Rubin's rules (Little, 1992). To obtain a consistent estimate for the gold standard, we always estimate $\hat{E}\{Y(c);\hat{\boldsymbol{\beta}},\hat{\psi}\}$ in (4.4) based on MI data, while $\hat{E}_{(\mathbf{S})}\{Y^*(c);\hat{\boldsymbol{\beta}}^*_{(\mathbf{S})},\hat{\psi}^*_{(\mathbf{S})}\}$ may be based on CC or MI data. We presume that problems of residual confounding pose no major concern here, because patient mix has been shown not to differ substantially across centers for Riksstroke (Varewyck et al., 2014, 2015).

We aim to find the subset of covariates that results in the smallest error $ER_1(\mathbf{S})$ or $ER_2(\mathbf{S})$. For these stochastic searches we only consider multiple imputation to handle missing values, because it has repeatedly been shown to be superior over a complete case analysis (van der Heijden et al., 2006; Little, 1992) although also the more challenging method in combination with variable selection (Wood et al., 2008). Our patient covariate cost is determined by the proportion of missing values: 1 (no missing values), 1.5 (up to 5% missing values), 2 (up to 50% missing values) or 3 (more than 50% missing values) and we constrain the total allowed cost to 7 (Table 4.3 in Appendix). Higher covariate costs then reflect the effort it would take to fill in the missing values after the data were collected. Evaluating all $2^{18} \approx 2.6 \times 10^5$ possible combinations is infeasible, because evaluating the error for one given subset is already computationally demanding and may take up to 6 minutes. Therefore we use a stochastic hill-climber with 10 restarts that each stop after 2 hours without improvement, although 9 hours are needed to evaluate all possible neighbors of a given subset. For this time-consuming search, we perform only one cross-validation split. When splitting the multiple imputed data, we fix the selected subjects over the imputations. We will use standardized

age and log-transformed NIHSS as we found a good fit for a loglinear effect of NIHSS on 30-day mortality risk (Figure 4.10 in Appendix).

The subset of patient characteristics with the smallest error at patient level is $\mathbf{S}_1$: stroketype, age, consciousness level and NIHSS, with total cost 6.5 and $\hat{ER}_1(\mathbf{S}_1) = 0.2787$. This subset includes NIHSS which has the largest covariate cost and one covariate which is measured prior to stroke, namely age. As this subset with the costly NIHSS is selected only once over the 10 restarts (Figure 4.7 in Appendix), it is likely that the search algorithm was trapped in a suboptimal solution for the other restarts or not given enough time to find a better neighbor. We found a different best subset of patient characteristics when minimizing the error on the directly standardized risks, $\mathbf{S}_2$: year of admission, patient's ADL-dependence, consciousness level and NIHSS, with total cost 7 and average error over the centers $\hat{ER}_2(\mathbf{S}_2) = 0.0161$ (Figure 4.8 in Appendix). Comparing the subsets $\mathbf{S}_1$ and $\mathbf{S}_2$, the predictors stroketype and age were replaced by year of admission and patient's ADL-dependence in the context of health care evaluations. Indeed, hospital performance is very likely to be confounded by the year of admission and by the degree of patients' dependence or thus how strongly patients rely on the provided care level.

As the best subsets $\mathbf{S}_1$ and $\mathbf{S}_2$ include both consciousness and NIHSS, which are different measures for the same patient's baseline disease severity, we now aim to investigate whether consciousness is a good surrogate for NIHSS.

4.4 Analytical Reflections on the Inclusion of Covariates

We compare the error change when including a covariate with missing values (e.g. NIHSS) versus a surrogate which is completely, but imprecisely measured (e.g. consciousness). These comparisons will be made for the prediction error (4.3) and the error on the directly standardized risk of a given center c (4.4); details on the calculations can be found in the Appendix, Section 4.A. We will consider the simple setting of a linear regression model for a continuous outcome Y, including two patient characteristics and two centers. Such model is equivalent

with a model including mean-centered covariates:

$$Y = \beta_1 L_1 + \beta_2 L_2 + \psi_1 I(C = 1) + \psi_2 I(C = 2) + \varepsilon, \tag{4.8}$$

where ε is a random variable with mean zero and variance σ_ε^2 conditional on (L_1, L_2, C) and $E(L_1) = E(L_2) = 0$. As before, using cross-validation, we randomly split the data into two complementary parts, $\{(Y_i; L_{1i}, L_{2i}, C_i), i = 1, \ldots, n\}$ and $\{(Y_i^*; L_{1i}^*, L_{2i}^*, C_i^*), i = 1, \ldots, n^*\}$, which are assumed to have the same distribution. Let L_1 be a surrogate for L_2, so that L_1 contains no other information on Y than what is available in L_2. More formally, we assume that $L_1 = L_2 + U$ with $E(U|L_2, C) = 0$ and that Y is conditionally independent of L_1 given L_2 and C, so that $\beta_1 = 0$ in the regression model (4.8).

4.4.1 A covariate with measurement error

First, we consider the working outcome regression model including only the surrogate L_1:

$$Y = \alpha_1 L_1 + \alpha_2 I(C = 1) + \alpha_3 I(C = 2) + \varepsilon_s, \tag{4.9}$$

where ε_s has mean zero and variance σ_s^2 conditional on (L_1, C). The residual variance of this regression is (Carroll et al., 2006):

$$\sigma_s^2 = \sigma_\varepsilon^2 + \frac{\beta_2^2 \sigma_u^2 \sigma_2^2}{\sigma_u^2 + \sigma_2^2}, \tag{4.10}$$

where $\sigma_u^2 = \text{Var}(U|L_2, C)$ and $\sigma_2^2 = \text{Var}(L_2|C)$. Let $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3)$ be the least squares estimates of the model parameters $(\alpha_1, \alpha_2, \alpha_3)$. As shown by equation (4.20) in the Appendix, the expected prediction error for a 'new' patient with characteristic L_1^* and treated at center C^*, is:

$$E\left[\left\{Y^* - (\hat{\alpha}_1 L_1^* + \hat{\alpha}_2 I(C^* = 1) + \hat{\alpha}_3 I(C^* = 2))\right\}^2\right] = \sigma_s^2 \left(1 + \frac{p}{n}\right), \tag{4.11}$$

where $p = 3$ is the number of model parameters. From (4.10) it is clear that σ_s^2 will never be smaller than σ_ε^2, which confirms the intuition that the residual variance increases by regressing Y on the 'less-informative' surrogate L_1 instead

of on L_1 and L_2. Moreover, σ_s^2 and thus the prediction error will increase for a stronger predictive value of L_2 on the outcome, as determined by β_2, or with larger measurement error.

For the standardized risks, estimation of the model parameters and the directly standardized risks are both based on $(Y^*; L_1^*, L_2^*, C^*)$. Therefore, notation will be simplified by omitting the star superscript in the notation below for the directly standardized risks. Under the assumption that the patient covariates are mean-centered, the directly standardized risk at center c is simply ψ_c, $(c = 1, 2)$. As shown by equation (4.22) in the Appendix, the expected error on the directly standardized risk for center 1 is then:

$$E\left\{(\hat{\psi}_1 - \psi_1)^2\right\} = \frac{\sigma_s^2}{n}\left\{\frac{1}{P(C=1)} + \frac{E(L_1|C=1)^2}{E(L_1^2)}\right\}. \tag{4.12}$$

If there is no difference in patient-mix across centers, then $E(L_1) = E(L_1|C=1) = E(L_1|C=2) = 0$, so that the error simplifies to $\sigma_s^2\{nP(C=1)\}^{-1}$, showing that larger centers will have a smaller error on their directly standardized risk, which corresponds to previous findings in Varewyck et al. (2015).

4.4.2 A covariate with missing values

Suppose now that we include the more precise, but incomplete measure L_2 instead of its surrogate L_1. Let $\mathbf{R} = (R_1, \ldots, R_n)^T$ where $R_j = 1$ if L_2 is observed and $R_j = 0$ if L_2 is missing for the j-th subject $(j = 1, \ldots, n)$. We assume that $R \perp\!\!\!\perp Y | L_1, L_2, C$ and perform a complete case analysis. Moreover, we assume that the covariate L_2 is mean-centered given the complete cases, $E(L_2|R=1) = 0$, and that (L_2^*, C^*) is a random sample from the complete cases. Let $(\hat{\beta}_2, \hat{\psi}_1, \hat{\psi}_2)$ be the least squares estimates of the model parameters $(\beta_2, \psi_1, \psi_2)$ in (4.8), based on the complete cases. If the distributions of (L_2, C) in the complete cases and (L_2^*, C^*) are equal, the expected prediction error is:

$$E\left[\left\{Y^* - (\hat{\beta}_2 L_2^* + \hat{\psi}_1 I(C^*=1) + \hat{\psi}_2 I(C^*=2))\right\}^2\right] = \sigma_\varepsilon^2\left\{1 + \frac{p}{nP(R=1)}\right\}, \tag{4.13}$$

where $p = 3$ is again the number of model parameters. It is clear that the prediction error for a complete case analysis with a model including only L_2 will be smaller than for the model including only the completely measured surrogate L_1 if,

$$\frac{\sigma_s^2}{\sigma_\varepsilon^2} > 1 + \frac{1 - P(R=1)}{P(R=1)\left(1 + \frac{n}{p}\right)}, \tag{4.14}$$

or e.g. if there are many observations in the data set for parameter estimation ($n \gg$), few missing values and thus $P(R=1) \approx 1$ or if the measurement error is large $\sigma_s^2 \gg \sigma_\varepsilon^2$. Let λ express which fraction of the variability in L_1 is due to measurement error in the sense that $\lambda = \text{Var}(U|C)\,\text{Var}(L_1|C)^{-1}$. Then it is shown by equation (4.23) in the Appendix that (4.14) is equivalent with:

$$\left(1 + \frac{n}{p}\right)\lambda \frac{R^2_{L_2|C}}{1 - R^2_{L_2|C}} > \text{odds}(R=0), \tag{4.15}$$

where $R^2_{L_2|C}$ is the coefficient of partial determination between Y and L_2 given C. This may help to decide for which percentage of missing values it is still beneficial to include the more informative variable L_2 when there is a completely measured surrogate L_1 available, as will be illustrated in Section 4.4.3.

When regressing Y on L_2, the directly standardized risk for center c ($c = 1,2$) is again expressed by ψ_c, under the additional assumption that the observed distribution of L_2 is the same as in the study population, $R \perp\!\!\!\perp L_2$. Then, the expected error on the directly standardized risk for center 1 is:

$$\begin{aligned} E\left\{(\hat{\psi}_1 - \psi_1)^2\right\} &= \frac{\sigma_\varepsilon^2}{n\,P(R=1)}\left\{\frac{1}{P(C=1|R=1)} + \frac{E(L_2|C=1,R=1)^2}{E(L_2^2|R=1)}\right\} \\ &= \frac{\sigma_\varepsilon^2}{n\,P(R=1)\,P(C=1|R=1)}, \end{aligned} \tag{4.16}$$

where the last equality is only justified if there is no difference in patient-mix across centers for the complete cases. Here again, the largest centers will have the smallest error on their directly standardized risk. Under the assumption of equal patient-mix across centers, the error on the directly standardized risk for

center 1 is smaller when regressing Y on L_2 instead of its surrogate L_1, if:

$$\frac{\sigma_s^2}{\sigma_\varepsilon^2} > \frac{P(C=1)}{P(R=1)P(C=1|R=1)}, \tag{4.17}$$

which is most probable if there are few missing values or the measurement error is large. As shown by equation (4.24) in the Appendix, the error on the averaged directly standardized risks, giving equal weight to each center, then becomes:

$$1+\lambda\frac{R^2_{L_2|C}}{1-R^2_{L_2|C}} > \frac{\sum_{c=1}^{m}\frac{1}{P(C=c|R=1)}}{P(R=1)\sum_{c=1}^{m}\frac{1}{P(C=c)}}. \tag{4.18}$$

If $R \perp\!\!\!\perp C$, the latter simplifies to:

$$\lambda\frac{R^2_{L_2|C}}{1-R^2_{L_2|C}} > \text{odds}(R=0), \tag{4.19}$$

which is less likely to be fulfilled than the similar criterion at patient level (4.15). So, even when the variance of the surrogate would be much larger than for L_2 (e.g. $\lambda = 50\%$) and this more precise measure would be hardly missing (e.g. odds($R = 0$) = 10%), the relative reduction in residual error by additionally including L_2 on top of C has to be substantial (e.g. $R^2_{L_2|C} = 20\%$) in order to favor the inclusion of L_2 instead of L_1 (see also equation (4.25) in Appendix).

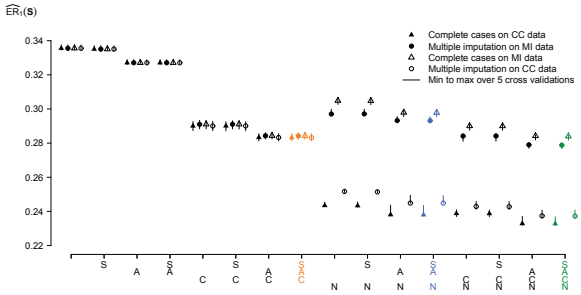
In conclusion, measurement errors have a direct adverse effect on the accurateness of predicted outcomes, so that in most settings the analytical results favored the inclusion of the most accurate measure instead of its surrogate, even when the former is often missing. This will however reduce for which patients predictions can be obtained. On the other hand, when the aim is to assess hospital performance, more observations will improve the accurateness of the estimated center effects, even when the included confounders are measured with some error. So, in this case missing values have a bigger impact than measurement error, and consequently the inclusion of the available surrogate was favored in most settings. Notice that these analytic results only hold for a complete case analysis and that a similar derivation for multiple imputed data is much more complicated.

4.4.3 Comparing the added value of consciousness and NIHSS

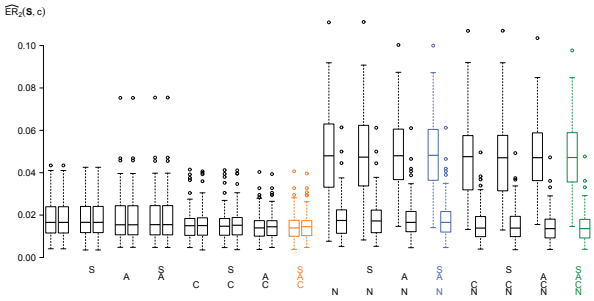
For Riksstroke, we aim to assess the benefit of NIHSS over consciousness as measure of baseline disease severity in the prognostic score via the error at patient or hospital level. We will study differences in the error analytically and by performing $V = 5$ cross-validations to evaluate the errors $ER_1(\mathbf{S})$ and $ER_2(\mathbf{S})$ for all possible subsets of the baseline covariates log NIHSS, consciousness, sex and age. The latter three patient characteristics are chosen because national reports on Riksstroke traditionally adjust the outcome for those (Stroke Board Team, 2011). We found that NIHSS values tend to be larger for drowsy or unconscious patients than for alert patients (Figure 4.9 in Appendix), suggesting that consciousness may be a surrogate for NIHSS.

Based on the analytical results in Sections 4.4.1 and 4.4.2, we can first get some intuition on whether to best include consciousness or NIHSS without explicitly estimating the error. We simplify the setting by excluding 1338 patients with missing consciousness level, which left a sample of 122 970 patients of which 41 798 ($\hat{P}(R = 1) = 33\%$) patients have observed NIHSS. Based on the data, $\sum_{c=1}^{m} 1/\hat{P}(C = c|R = 1) = 36\,726$, and $\sum_{c=1}^{m} 1/\hat{P}(C = c) = 11\,790$. With $p = m + 1 = 81$ and $n = 122\,970 \times 0.8$, where 0.8 is the proportion of data used for parameter estimation following cross-validation, equation (4.14) states that the model including NIHSS will have a smaller prediction error than the model including consciousness if $\hat{\sigma}_s^2 > (1 + 0.0013)\,\hat{\sigma}_\varepsilon^2$, which is highly likely. In contrast, for the directly standardized risk, this is most unlikely, unless $\hat{\sigma}_s^2 > 3 \times 3.12\,\hat{\sigma}_\varepsilon^2$.

For the MI data we found a smaller error, both at patient and hospital level, when including consciousness rather than NIHSS, in addition to sex and age (Figure 4.3). This is surprising, because NIHSS is a more comprehensive measure for baseline disease severity, but unconsciousness appears to be the stronger predictor of death following a univariate regression model (Table 4.3 in Appendix). For CC the smallest prediction error is obtained when including NIHSS, which was often missing, instead of consciousness. However, this is a misleading result as the error is evaluated on the selective subset of patients with observed NIHSS. Therefore in Figure 4.3a we additionally show the estimated prediction errors for fitting the models on CC while evaluating the prediction error on MI data and



(a) Median $E\hat{R}_1(\mathbf{S})$ (patient level) with range over the 5 cross-validations.



(b) Distribution of $E\hat{R}_2(\mathbf{S}, c)$ over all centers, averaged over 5 cross-validations; in pairs of CC analysis on the left and MI analysis on the right.

Figure 4.3: Estimated errors for null model or combination of patient characteristics sex (S), age (A), consciousness (C) and NIHSS (N). We highlight three covariate combinations that are of special interest for comparison.

found a smaller prediction error for the model including consciousness compared to the model including NIHSS, besides sex and age. The error at hospital level is remarkably large for models including NIHSS under a CC analysis, which was already clear from the analytic results but may also partly be due to estimating the gold standard as the full population average, hence based on MI data. In either case, the smallest errors were obtained for the model including sex, age, consciousness and NIHSS, although the additional error reduction compared to the model excluding NIHSS is negligible. For this reason and to avoid selection bias, we prefer not to include NIHSS on top of consciousness.

4.5 Discussion

Electronic health registers have expanded enormously in the last years. Their size and measurement cost are expected to increase even further e.g. when biomarkers are more routinely measured. For individual patient management and treatment decisions, physicians are interested in those risk factors that best predict mortality while limiting the overall workload and measurement cost. When initial measurements for disease severity point towards an aggressive cancer variant, one may additionally perform more expensive measurements that can guide finding the most effective targeted treatment (Spahn et al., 2015). In that case, values are missing for a selective group of patients and we have assessed whether the added predictive value then weighs against the average additional measurement cost. When the aim is to compare hospital performance, it is important to adjust for differences in patient mix that induce confounding and thus the interest lies primarily in those predictors of mortality that are also differentially distributed between centers. Here again, the number of covariates that can be measured may be restricted by the budget. In this paper, we suggested a simple and efficient approach to determine an appropriate subset of covariates to include for each of the settings. We illustrated results on the RAND data, which included a large number of patient characteristics, and the Swedish register Riksstroke, that allowed for comparisons of hospital performance. For the latter we focused on direct standardization, but our approach can similarly be applied to indirect standardization such as the excess risk, which contrasts the quality outcome in each center with what is expected if its patients were randomly assigned over the level of care across all centers (Goetghebeur et al., 2011).

Both the definition of the error function and the options for the search algorithm are very flexible. In this paper, we used a strict maximum budget, so that we rejected any model with total cost larger than the upper bound. Rather than spending any money - as long as it stays below the maximum budget - one may require a minimum benefit per 'dollar' invested. The search methods we used can indeed incorporate a penalty that accepts 1 unit cost increase provided the error is decreased by at least x units. Another option of the search algorithms is to increase the benefit per dollar required when the current subset's

cost approaches the budget constraint. In the literature, costs and prediction gain are sometimes placed on a common scale, trading one against the other, and optimization then occurs on that scale (Fouskakis and Draper, 2008).

The definition of the error function may vary depending on the specific context and does not fundamentally affect the search algorithm strategies. We considered one specific error function at the patient level (4.3) and one at the hospital level (4.4), which all give equal weight to 'positive' or 'negative' errors. The situation is not necessarily symmetric however. For example, when the difference between 'true' and estimated standardized risk is positive, we underestimate the risk at that hospital and high mortality may be masked. A different weighting factor for positive and negative differences can be obtained by a simple change in the definition of the error function. Although it is unclear how the errors on the directly standardized risks are best averaged over hospitals, we gave equal weight to each hospital but alternatively weights proportional to the number of registered patients could be given and one could minimize the mean absolute rather than squared error. Our regression model contained main effects only. In a second step, one may investigate interaction effects among those main effects. This step would no longer be constrained by the budget for registration.

Stochastic search algorithms were used to approximate the best subset selection within a reasonable time. The basic stochastic hill-climber performed very well as long as we allowed a sufficiently long computation time and restarted the procedure with a number of randomly chosen initial subsets. However, the stopping criterion is rather arbitrary and highly dependent on the specific setting, so that better solutions may be missed when it is set too strict. Therefore it may be interesting to monitor convergence during the search and stop the algorithm as soon as convergence is reached, for example by examining the error reduction relative to the amount of search time to find this better subset. To reduce the total search time, one may also consider to first perform a generalized LASSO regression, weighting covariates by their cost but not restricting the total cost. Given the best subset found via this procedure, the subset can be further reduced to meet the total cost constraint via the stochastic hill-climber.

When performing a complete case analysis the end user must be warned for misleading results when covariates are selected based on the estimated error.

This because including a covariate which is often missing, changes the subset of patients on which the error can be evaluated. A smaller prediction error may then not necessarily yield a better prediction model on the full study population. In practice, individual predictions for patients with missing covariate values will be based on their available patient characteristics and a separate model will be fit for each set of observed patient characteristics. Unfortunately, for the standardized risks this is no longer possible as estimation and standardization have to be performed on the same set of subjects. Comparing the predictive value of NIHSS and consciousness surprisingly favored the latter. Moreover, when including both in a model with sex and age, the added value of NIHSS on top of consciousness was limited. So, we have several reasons to prefer consciousness over NIHSS here: NIHSS is more absent, which is an important deficiency for a predictor and it also forced us to make assumptions on the reasons for missingness. Then, even if we assumed that values for NIHSS were missing at random, it did not beat the predictive value of consciousness. When NIHSS would get measured for more patients in the future, this may be reassessed before drawing general conclusions.

In all, we believe that the flexibility in defining the error function as well as the options of the search algorithms form a major advantage over other variable selection methods. Moreover, they can be combined with a cost-efficient generalized LASSO to further reduce the computation time. Some R packages implement these or similar search algorithms, such as in 'caret' (Kuhn, 2008) but lack the wealth of flexible options of the JAMES framework (De Beukelaer et al., 2015).

Acknowledgement

The authors are grateful to David Draper and Dimitris Fouskakis for providing data from the RAND DRG Quality of Care study.

4.A Analytical Reflections on the Inclusion of Covariates

In Section 4.4 we compare the error change when including a covariate with missing values versus its surrogate which is completely, but imprecisely measured. Here, we provide extra results on the calculations, building on the assumptions and notation introduced in Section 4.4. The outcome regression model in (4.8) can be written as $Y = \mathbf{X}\boldsymbol{\gamma} + \varepsilon$, where $\boldsymbol{\gamma} = (\beta_1, \beta_2, \psi_1, \psi_2)^T$ and $\mathbf{X} = (L_1, L_2, I(C=1), I(C=2))$. Let $\mathbf{X_n}$ be the $(n \times 4)$ design matrix for n individuals, with row $i : (L_{1i}, L_{2i}, I(C_i=1), I(C_i=2))$, $i = 1, \ldots, n$. Least squares estimation for $\boldsymbol{\gamma}$ then results in:

$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = (\mathbf{X_n}^T\mathbf{X_n})^{-1}\mathbf{X_n}^T(\mathbf{Y} - \mathbf{X_n}\boldsymbol{\gamma}) = (\mathbf{X_n}^T\mathbf{X_n})^{-1}\mathbf{X_n}^T\varepsilon,$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$.

In general, the expected prediction error for a 'new' patient with characteristics $\mathbf{X}^* = (L_1^*, L_2^*, I(C^*=1), I(C^*=2))$, is:

$$\begin{aligned}
E\left\{(Y^* - \mathbf{X}^*\hat{\boldsymbol{\gamma}})^2\right\} &= E\left\{(Y^* - \mathbf{X}^*\boldsymbol{\gamma})^2\right\} + E\left\{(\mathbf{X}^*(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}))^2\right\} \\
&= \sigma_\varepsilon^2 + E\left(\mathbf{X}^*(\mathbf{X_n}^T\mathbf{X_n})^{-1}\mathbf{X_n}^T\varepsilon\varepsilon^T\mathbf{X_n}(\mathbf{X_n}^T\mathbf{X_n})^{-1}\mathbf{X}^{*T}\right) \\
&= \sigma_\varepsilon^2 + \sigma_\varepsilon^2 E\left(\mathbf{X}^* E(\mathbf{X_n}^T\mathbf{X_n})^{-1}\mathbf{X}^{*T}\right) \\
&= \sigma_\varepsilon^2\left[1 + n^{-1}E\left\{\mathbf{X}^* E(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{*T}\right\}\right] \\
&= \sigma_\varepsilon^2\left\{1 + n^{-1}E\left(\mathbf{X}^T\mathbf{X}\right)^{-1}E\left(\mathbf{X}^*\mathbf{X}^{*T}\right)\right\} \\
&= \sigma_\varepsilon^2\left(1 + \frac{p}{n}\right), \qquad (4.20)
\end{aligned}$$

where p is the number of model parameters and the last equality is justified under the assumption that $(Y; L_1, L_2, C)$ and $(Y^*; L_1^*, L_2^*, C^*)$ are equally distributed. In the case of mean-centered patient covariates, the directly standardized risk at center c is simply ψ_c, $(c = 1, 2)$. The expected error for center 1 is then:

$$\begin{aligned}
E\left\{(\hat{\psi}_1 - \psi_1)^2\right\} &= E\left\{(0,0,1,0)(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T(0,0,1,0)^T\right\} \\
&= E\left\{(0,0,1,0)(\mathbf{X_n}^T\mathbf{X_n})^{-1}\mathbf{X_n}^T\varepsilon\varepsilon^T\mathbf{X_n}(\mathbf{X_n}^T\mathbf{X_n})^{-1}(0,0,1,0)^T\right\}
\end{aligned}$$

$$= \frac{\sigma_\varepsilon^2}{n}(0,0,1,0)E(\mathbf{X}^T\mathbf{X})^{-1}(0,0,1,0)^T. \tag{4.21}$$

For the first setting, the outcome regression model includes only L_1, which is the surrogate for L_2. In this case, the expected error on the directly standardized risk for center 1 in (4.12) is calculated as:

$$\begin{aligned}
&E\left\{(\hat{\psi}_1-\psi_1)^2\right\}\\
&= \frac{\sigma_s^2}{n}(0,1,0)\\
&\quad \begin{pmatrix} E(L_1^2) & P(C=1)E(L_1|C=1) & P(C=2)E(L_1|C=2)\\ P(C=1)E(L_1|C=1) & P(C=1) & 0\\ P(C=2)E(L_1|C=2) & 0 & P(C=2)\end{pmatrix}^{-1}\\
&\quad (0,1,0)^T\\
&= \frac{\sigma_s^2}{n}\left\{\frac{1}{P(C=1)}+\frac{E(L_1|C=1)^2}{E(L_1^2)}\right\}.
\end{aligned} \tag{4.22}$$

When L_2 is incompletely measured, we assume that $R \perp\!\!\!\perp Y|\mathbf{X}$ and perform a complete case analysis, so that the following identities hold,

$$E(Y|R=1,\mathbf{X}) = E(Y|\mathbf{X}) \quad \text{and} \quad \text{Var}(Y|R=1,\mathbf{X}) = \text{Var}(Y|\mathbf{X}) = \sigma_\varepsilon^2.$$

Let diag($\mathbf{R}$) be an $(n \times n)$ diagonal matrix with diagonal $\mathbf{R}$. Using complete case analysis, we estimate γ by solving the equations $\mathbf{X_n}^T \text{diag}(\mathbf{R})(\mathbf{Y}-\mathbf{X_n}\hat{\gamma}) = \mathbf{0}$, so that $\hat{\gamma}-\gamma = (\mathbf{X_n}^T \text{diag}(\mathbf{R})\mathbf{X_n})^{-1}\mathbf{X_n}^T \text{diag}(\mathbf{R})\varepsilon$. We first show that:

$$\begin{aligned}
E\left\{\mathbf{X_n}^T \text{diag}(\mathbf{R})\,\mathbf{X_n}\right\} &= E(\mathbf{X_n}^T\mathbf{X_n}|R=1)P(R=1)\\
&= nP(R=1)E\left(\mathbf{X}^T\mathbf{X}|R=1\right).
\end{aligned}$$

Using this result, the expected prediction error in (4.13) and the expected error on the directly standardized risk for center 1 in (4.16) can be calculated as before.

The prediction error for a complete case analysis with a model including only L_2 will be smaller than for the model including only the completely measured

surrogate L_1, if:

$$\begin{aligned}
\frac{\sigma_s^2}{\sigma_\varepsilon^2} &> \left\{1+\frac{p}{nP(R=1)}\right\}\left(1+\frac{p}{n}\right)^{-1} \\
1+\frac{\beta_2^2\sigma_u^2\sigma_2^2}{\sigma_\varepsilon^2(\sigma_u^2+\sigma_2^2)} &> 1+\frac{1-P(R=1)}{P(R=1)(n/p+1)} \\
1+\lambda\beta_2^2\frac{\sigma_2^2}{\sigma_\varepsilon^2} &> 1+\frac{\text{odds}(R=0)}{n/p+1} \\
(1+n/p)\,\lambda\frac{R^2_{L_2|\cdot}}{1-R^2_{L_2|\cdot}} &> \text{odds}(R=0). \qquad (4.23)
\end{aligned}$$

where $R^2_{L_2|\cdot}$ is the coefficient of partial determination between Y and L_2 given the other covariates in the outcome regression model (4.8). The average error on the directly standardized risks is smaller when regressing Y on L_2 instead of the surrogate L_1, if:

$$\begin{aligned}
\sum_{c=1}^{m}\frac{\sigma_\varepsilon^2}{nP(R=1)P(C=c|R=1)} &< \sum_{c=1}^{m}\frac{\sigma_s^2}{nP(C=c)} \\
\frac{\sigma_s^2}{\sigma_\varepsilon^2} &> \frac{\sum_{c=1}^m\frac{1}{P(C=c|R=1)}}{P(R=1)\sum_{c=1}^m\frac{1}{P(C=c)}} \\
1+\lambda\beta_2^2\frac{\sigma_2^2}{\sigma_\varepsilon^2} &> \frac{\sum_{c=1}^m\frac{1}{P(C=c|R=1)}}{P(R=1)\sum_{c=1}^m\frac{1}{P(C=c)}} \\
1+\lambda\frac{R^2_{L_2|\cdot}}{1-R^2_{L_2|\cdot}} &> \frac{\sum_{c=1}^m\frac{1}{P(C=c|R=1)}}{P(R=1)\sum_{c=1}^m\frac{1}{P(C=c)}}. \qquad (4.24)
\end{aligned}$$

If $R \perp\!\!\!\perp C$, the latter simplifies to:

$$\begin{aligned}
\lambda\frac{R^2_{L_2|\cdot}}{1-R^2_{L_2|\cdot}} &> \text{odds}(R=0) \\
\frac{1}{1-R^2_{L_2|\cdot}} &> \frac{\text{odds}(R=0)}{\lambda}+1 \\
R^2_{L_2|\cdot} &> \frac{\text{odds}(R=0)}{\lambda+\text{odds}(R=0)}. \qquad (4.25)
\end{aligned}$$

4.B Additional Figures and Tables

Variable		
Index	Name	Cost
1	**Systolic blood pressure score**	0.5
2	Age	0.5
3	**Blood urea nitrogen**	1.5
4	APACHE II coma score	2.5
5	**Shortness of breath day 1**	1.0
6	**Serum albumin score**	1.5
7	Respiratory distress	1.0
8	Septic complications	3.0
9	Prior respiratory failure	2.0
10	Recently hospitalized	2.0
11	Racbilateral process score	1.5
12	Initial temperature	0.5
13	Heart rate day 1	0.5
14	Chest pain day 1	0.5
15	Cardiomegaly score	1.5
16	Plural effusion score	1.5
17	Chest X-ray congestive heart failure score	2.5
18	Ambulatory score	2.5
19	Endocarditis at admission	1.5
20	Creatine phosphokinase score	2.0
21	**Prior antibiotics**	0.5
22	Prior interstitial lung disease	0.5
23	Home oxygen use	1.0
24	Prior pneumonectomy	0.5
25	Prior tracheostomy	0.5
26	Prior aminophylline score	0.5
27	Haematologic history score	1.5
28	Cancer score	1.5
29	APACHE heart rate score	1.5
30	Corodaker score	1.0
31	Disease of thorax	1.0
32	Multiple myeloma	0.5
33	Immunocompromised	0.5
34	Residence score	1.0
35	Hepatobiliary history	0.5
36	Renal history score	1.0
37	**APACHE respiratory rate score**	1.0
38	New lung score	1.0
39	Comorbid aspiration score	0.5
40	APACHE sodium score	2.0
41	APACHE haematocrit score	1.5
42	APACHE white blood cell score	1.5
43	APACHE oxygenation score	1.5
44	Cardiovascular accident score	1.0

45	APACHE potassium score	1.0
46	**Admission systolic blood pressure**	0.5
47	Congestive heart failure chest X-ray score	2.5
48	Total APACHE II score	10.0
49	Respiratory rate day 1	0.5
50	Diastolic blood pressure day 1	0.5
51	Confusion day 1	0.5
52	Pulmonary vascular congestion score	0.5
53	APACHE venous bicarbonate score	1.5
54	**Pulmonary of oedema score**	0.5
55	Sum of congestive heart failure components	5.5
56	Influenza score	0.5
57	**Arrest in emergency room score**	0.5
58	Biliribin score	1.5
59	Positive blood culture	0.5
60	Positive urine culture	0.5
61	Wheezing at admission	0.5
62	Body system count	2.5
63	Morbid prior chronic obstructive pulmonary disease score	0.5
64	Morbid pulmonary hospitalization score	0.5
65	Comorbid cirrhosis score	0.5
66	Comorbid congestive heart failure score	0.5
67	Comorbid arrhythmias score	0.5
68	**Comorbid smoking score**	0.5
69	Comorbid alcoholism score	0.5
70	**APACHE acidity score**	1.0
71	Comorbid nasogastric tubes score	0.5
72	Comorbid steroids score	0.5
73	Morbid + comorbid score	7.5
74	Cardiac history score	0.5
75	**Neurologic history score**	0.5
76	Oncologic history score	0.5
77	Immunologic history score	0.5
78	**Musculoskeletal score**	0.5
79	APACHE temperature score	1.0
80	APACHE mean blood pressure score	1.0
81	APACHE creatinine score	1.0
82	Diagnoses score	1.0
83	Sex of patient	0.5

Table 4.2: Full set of 83 variables for the RAND data set, together with their covariate cost per patient. In bold we indicate the variables that minimized the error on patient's predicted 30-day mortality, using the basic stochastic hill-climber and the parallel tempering algorithm, given a cost constraint of 10.

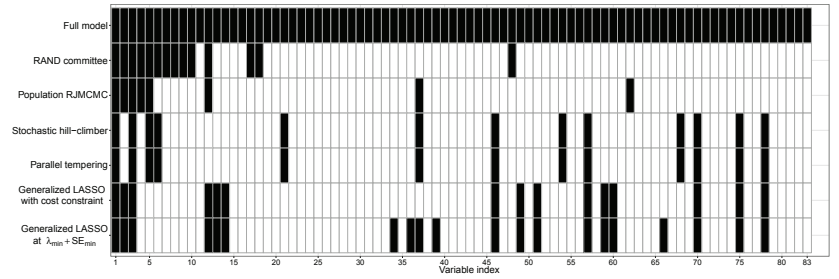


Figure 4.4: The selected subsets of covariates for each of the variable selection methods on the RAND data.

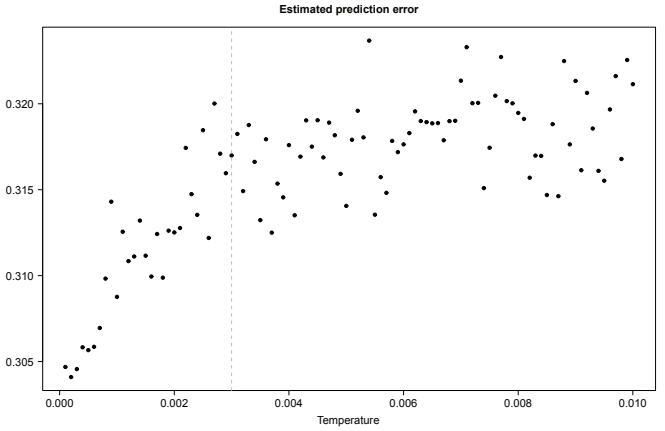
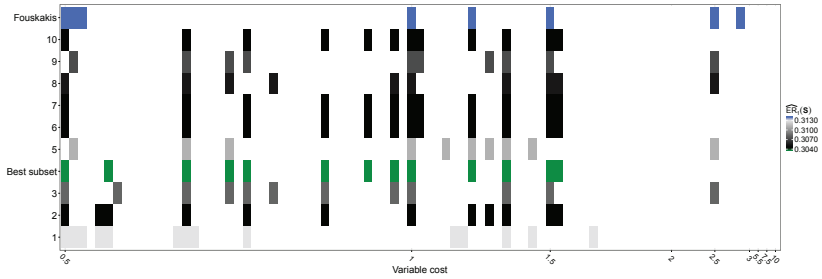
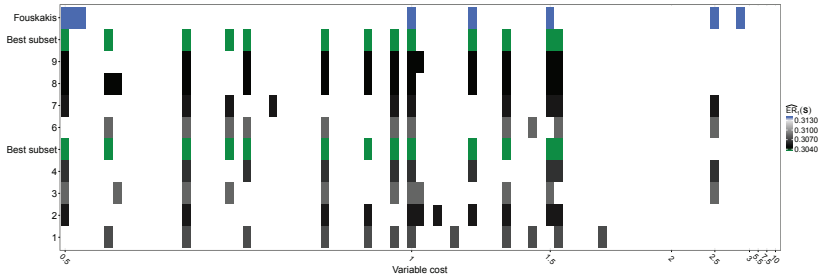


Figure 4.5: The estimated prediction error for the final best subset when performing 100 Metropolis searches with given temperature between 0 and 0.010. The maximum temperature for the parallel tempering algorithm is then set at 0.003.



(a) Stochastic hill-climber



(b) Parallel Tempering

Figure 4.6: For 10 restarts we report the selected subset of patient characteristics for the RAND data and compare the selected subset with the one found in Fouskakis et al. (2009).

	Prevalence (%)	Missing (%)	Cost	Univariate analysis Odds ratio	*p*-value
Male *	50.9	0	1	1.40	< 0.001
Age (in years) * (Mean & sd)	75.3 (12.4)	0	1	1.06	< 0.001
Consciousness at admission		1.1	1.5		< 0.001
(Alert)	82.6				
Drowsy	12.1			8.6	
Unconscious	5.3			38.71	
NIHSS (Mean & sd)	7.1 (8.8)	66.2	3	1.09	0.018
log NIHSS (Mean & sd)	1.6 (1.0)	66.2	3	2.92	0.052
p-ADL dependence *	10.0	1.8	1.5	3.98	< 0.001
Institutional living *	8.5	0.5	1.5	4.41	< 0.001
Living alone *	49.8	0.7	1.5	1.85	< 0.001
Atrial fibrilation *	27.2	1	1.5	2.11	< 0.001
Diabetes *	19.6	0.7	1.5	1.09	< 0.001
Trt for high blood pressure *	57.4	1.2	1.5	1.13	< 0.001
Current smoker *	14.8	9.8	2	0.54	< 0.001
CT scan	98.4	0.2	1.5	0.19	< 0.001
Thrombolysis	5.3	0.7	1.5	0.57	< 0.001
Stroke subtype		0	1		< 0.001
(Intracerebral haemorrhage (I61))	11.8				
Cerebral infarction (I63)	85.7			0.29	
Unspecified stroke (I64)	2.5			0.96	
Education *		3	1.5		< 0.001
(Primary)	49.9				
Secondary	35.1			0.68	
University	14.9			0.55	
Country of birth *		0.8	1.5		< 0.001
Sweden	88.1				
Other Nordic	5.4			0.83	
Other Europe	4.4			0.87	
Other	2.1			0.62	
Adjusted yearly income (in 100 SEK) *		0.8	1.5		< 0.001
(< 861)	9.3				
861 to 1330	31.1			0.97	
1330 to 2490	45.6			0.61	
> 2490	14.0			1.30	
Year of admission		0	1		0.012
(2007)	16.4				
2008	16.6			1.03	
2009	16.7			1.02	
2010	17.1			1.01	
2011	16.8			0.99	
2012	16.4			0.93	
30-day mortality	13.1	0	-	-	-
Distribution of 30-day mortality risk (%) over		CC (based on SACN model)		MI	
Centers (Mean and range)		8.0	0.0 to 30.8	12.8	6.7 to 19.1
Consciousness level: Alert		4.4		6.1	
Drowsy		28.8		35.8	
Unconscious		58.5		71.6	
NIHSS (quartiles): 0 to 1		1.3		2.8	
2 to 4		2.5		6.2	
5 to 9		5.8		12.3	
10 to 42		24.9		33.2	

*Table 4.3: Patient characteristics and outcome in Riksstroke, indicating * if measured prior to stroke; reference category for regression model between brackets. Results are averaged over the 5 imputed data sets for Riksstroke, except for CC results and missing (%). CC are the complete cases considering sex (S), age (A), consciousness (C) and NIHSS (N) and MI are multiple imputed data.*

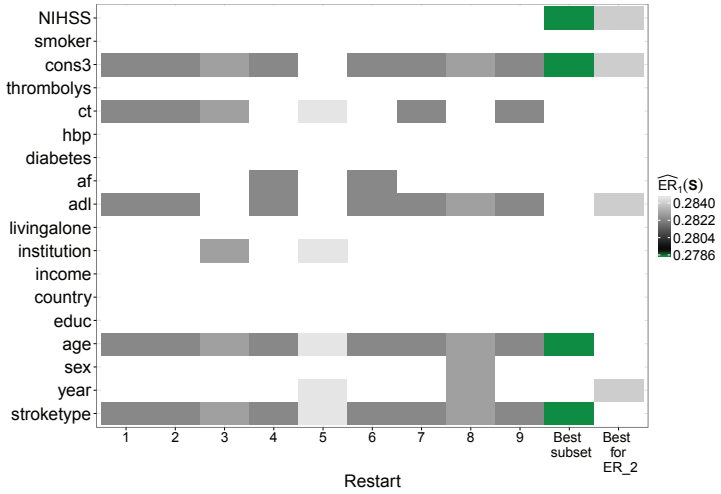


Figure 4.7: The selected subset of patient characteristics when minimizing the error at patient level $\hat{ER}_1(\mathbf{S})$ following stochastic hill-climber on multiple imputed data for Riksstroke. The longest computation time for the 10 independent restarts is 5.6 hours.

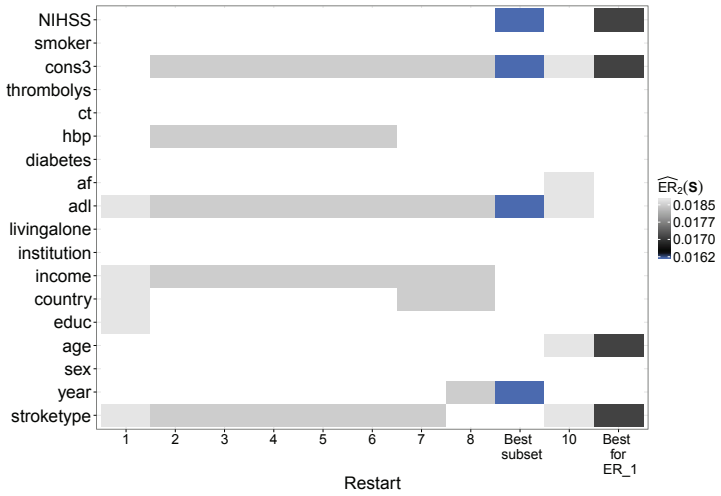


Figure 4.8: The selected subset of patient characteristics when minimizing the error at hospital level $\hat{ER}_2(\mathbf{S})$ following stochastic hill-climber on multiple imputed data for Riksstroke. The longest computation time for the 10 independent restarts is 7.1 hours.

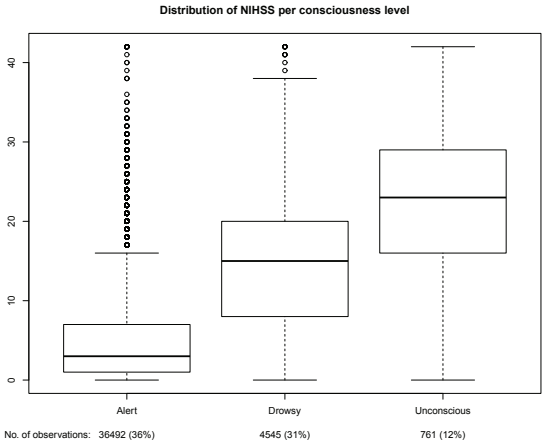


Figure 4.9: Measured NIHSS values over consciousness levels (restricted to patients with both observed), at the bottom indicating the number (and percentage) of patients with measured NIHSS values.

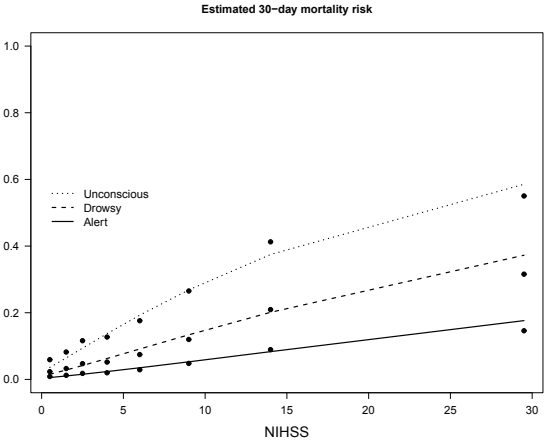


Figure 4.10: 30-day mortality risk in function of NIHSS per consciousness level for male patients with mean age treated at the reference hospital. Dots are estimated risks based on a model including 10% *percentile categories for NIHSS and lines assume a loglinear effect of NIHSS.*

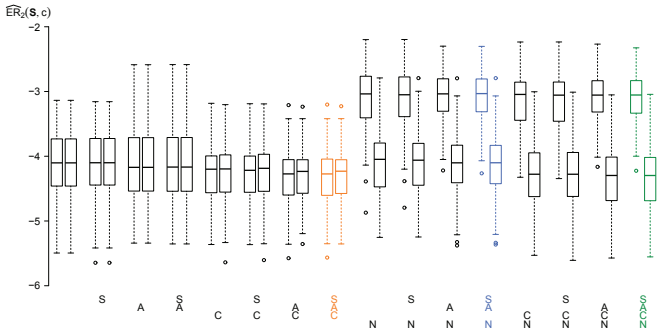


Figure 4.11: Estimated errors per combination of patient characteristics sex (S), age (A), consciousness (C) and NIHSS (N). Distribution of $\widehat{ER}_2(\mathbf{S}, c)$ over all centers, averaged over 5 cross-validations - log scale; in pairs of CC analysis on the left and MI analysis on the right.

CHAPTER 5

The R package RiskStandard

5.1 Introduction

This R package implements statistical methods for benchmarking clinical care centers based on a binary quality indicator such as 30-day mortality. For each center we provide directly or indirectly standardized risks based on fixed center effects in a logistic regression model that also incorporates patient-specific baseline covariates to adjust for differential case-mix. The user can choose to apply the Firth correction (Firth, 1993) to the logistic outcome model to maintain convergence in the presence of very small centers.

The package includes three example datasets: 'smallCaseMix', with small differences in patient mix across centers, 'largeCaseMix' with large differences in patient mix across centers, and 'largeCaseMix_missing' which is based on largeCaseMix but where the consciousness level is missing for some patients. Input data must contain for each patient (1) patient-specific covariates to adjust for in the analysis e.g. age, baseline disease severity, (2) a hospital code where the patient was treated and (3) a binary quality outcome e.g. 30-day mortality. In this document we will illustrate how center performance can be assessed using

the implemented R-functions.

We refer the user to Varewyck et al. (2014) for the theory behind the implemented R-functions. The R package is freely available at `www.cvstat.ugent.be`.

5.2 Implemented R-Functions

In this section we illustrate the two summary functions and three plot functions of the package. First, install and load the package 'RiskStandard':

```
> install.packages("./RiskStandard_0.0.6.tar.gz",
                   repos = NULL, type = "source")

> library(RiskStandard)
```


The dataset largeCaseMix contains for the $n = 50\,000$ patients treated in one of the $m = 50$ centers:

- patient-specific covariates: age (continuous), sex (binary) and consciousness level at admission (1=alert, 2=drowsy, 3=unconscious)
- hospital code: center (1 to m)
- binary quality outcome: 30-day mortality (0=alive, 1=dead)

Before assessing center performance, we recommend to make some descriptives of the dataset to get an impression of the distribution of patient characteristics across centers (Figure 5.1).

```
> str(largeCaseMix)

'data.frame':        50000 obs. of  5 variables:
 $ age    : int  65 74 64 69 76 87 61 75 51 71 ...
 $ sex    : int  0 0 0 1 0 0 0 1 1 1 ...
 $ cons   : Factor w/ 3 levels "1","2","3": 1 1 1 1 2 1 1 1 1 1 ...
 $ center : int  25 50 24 1 25 24 16 24 50 32 ...
 $ outcome: int  0 0 1 0 1 0 0 0 0 0 ...

> m <- length(unique(largeCaseMix$center))
> n <- dim(largeCaseMix)[1]
> centerSize <- as.vector(table(largeCaseMix$center))
```

```
> layout(matrix(1:4))
> # Age
> with(largeCaseMix,
      plot(1:m, tapply(age, center, mean), pch = 19,
           cex = centerSize/n*m, cex.lab = 1.2, ylim = c(60,80),
           xlab = "Center",  ylab = "", main = "Mean age per center"))
> with(largeCaseMix,
      boxplot(age ~ center, xlab = "Center", ylab = "Age distribution",
      cex.lab = 1.2))
> # Sex
> with(largeCaseMix,
      plot(1:m, tapply(sex, center, mean), pch = 19, cex=centerSize/n*m,
           cex.lab = 1.2, ylim = c(0,1), xlab = 'Center', ylab = "",
           main= "Percentage women per center"))
> # Consciousness
> with(largeCaseMix,
      plot(1:m, tapply((cons==1), center, mean), pch = 21,
           cex = centerSize/n*m, cex.lab = 1.2, ylim = c(0,1.1),
           xlab='Center', ylab = "",
           main="Distribution of consciousness level per center"))
> with(largeCaseMix,
      points(c(1:m), tapply((cons %in% c(1,2)), center, mean), pch = 19,
             cex = centerSize/n*m))
> legend("bottomleft", pch = c(21,19), bty='n',
        legend = c("Alert", "Alert or drowsy"), cex=1.2)
```

5.2.1 standardizeRisks()

This function estimates the standardized mortality risks. Necessary parameters are:

- patientCovariates: design matrix for the patient-specific covariates. Please make sure that categorical covariates are passed as factor, otherwise a linear effect will be assumed in the fitted outcome model (p covariate values for each of the n patients, giving an $n \times p$ data frame)
- center: hospital code (1 value out of m for each of the n patients)
- Y: binary quality outcome (0 or 1 for each of the n patients)

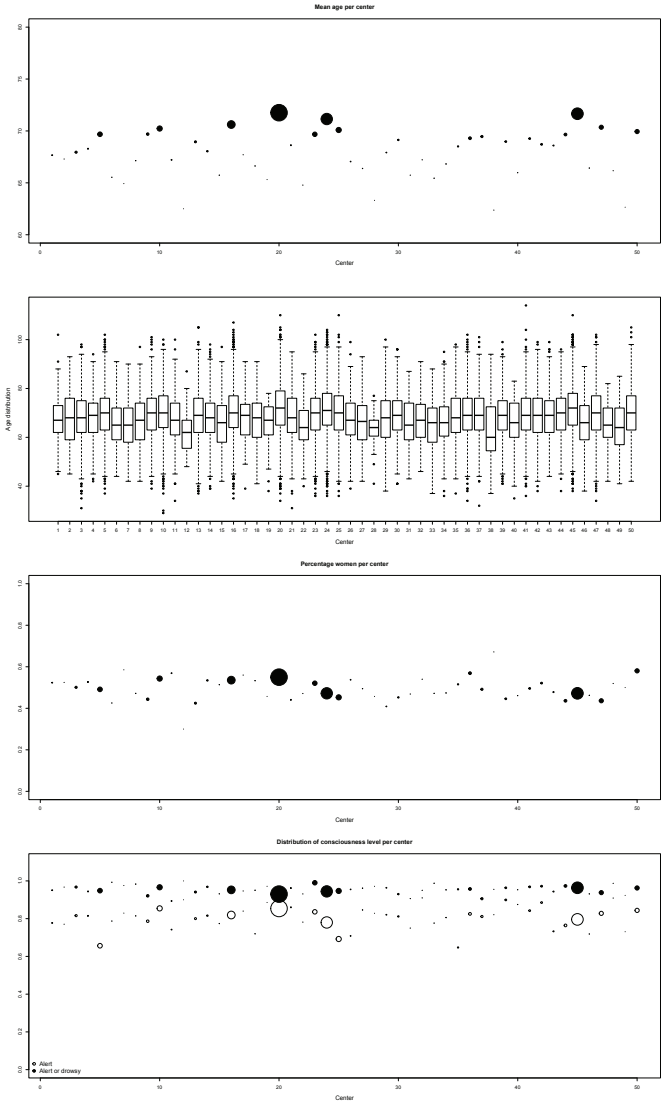


Figure 5.1: Descriptive plots for patient case-mix across centers. The size of the plot symbol is proportional to the center size.

The input for the parameter 'center' can be a character vector with the hospital names which will automatically be used in the output. The other function parameters have default values, but can be changed by the user (see documentation in R). By default, a Firth corrected outcome regression model is fitted. By default, no summary of the fitted model is printed, but it can be asked by adding the argument `trace = TRUE`. This summary can be useful e.g. to check whether each of the covariates was passed in the correct format.

```
> indirectRisks <- standardizeRisks(
   patientCovariates = largeCaseMix[,c('age','sex','cons')],
   center = largeCaseMix[,'center'],
   Y = largeCaseMix[,'outcome'])
> head(indirectRisks)

  centerName centerSize standardizedRisk varStandardizedRisk     lowerCI
1          1        386     -0.023367272        0.0003162722 -0.05822335
2          2         61     -0.007509684        0.0019252936 -0.09350931
3          3        984     -0.016308775        0.0001421959 -0.03968055
4          4        448     -0.075252885        0.0002203354 -0.10434600
5          5       2111      0.033703502        0.0001061479  0.01351036
6          6        141     -0.032481875        0.0006909590 -0.08400168
       upperCI observedRisk
1  0.011488811    0.1658031
2  0.078489944    0.1803279
3  0.007062996    0.1636179
4 -0.046159770    0.1093750
5  0.053896643    0.2676457
6  0.019037931    0.1347518
```

Similarly for direct standardization:

```
> directRisks <- standardizeRisks(
   patientCovariates = largeCaseMix[,c('age','sex','cons')],
   center = largeCaseMix[,'center'],
   Y = largeCaseMix[,'outcome'],
   method='direct')
```

When some patients have missing values for a categorical patient covariate such as consciousness level, we offer two ways to handle the missingness. The option `missing='completeCase'` (default) performs a complete case anal-

ysis, excluding all patients who have missing consciousness. The option `missing='dummyCategory'` adds a separate category to the fitted outcome model, allowing for a missing value effect. When some patients have missing values for a continuous patient covariate such as age, the function will by default perform a complete case analysis, excluding all patients who have missing age. Alternatively, multiple imputation can be considered to handle missingness. Although this method is currently not implemented in the standardizeRisks() function, the user can pass each of the imputed datasets separately to the function and afterward average the estimated standardized risks over the different imputations. The variance on the standardized risks can then be obtained by combining the within and between imputation variance as explained in Schafer (1999). The number of observations (n) that was used for the analysis can be extracted as an attribute from the function.


```
> indirectRisks2 <- standardizeRisks(
   patientCovariates = largeCaseMix_missing[,c('age','sex','cons')],
   center = largeCaseMix_missing[,'center'],
   Y = largeCaseMix_missing[,'outcome'],
   method='indirect', missing='completeCase')
> attr(indirectRisks2, "n")

[1] 38345
```

5.2.2 labelCenters()

The output from the `standardizeRisks()` function can then be used to classify the centers as having 'low', 'accepted' or 'high' mortality risk.

```
> labeledCenters <- labelCenters(standardizedRisks = indirectRisks)
> head(labeledCenters)

  centerName centerLabel     lowerCI      upperCI
1          1           A -0.03536243 -0.011372117
2          2           A -0.03710506  0.022085690
3          3           A -0.02435179 -0.008265759
4          4           L -0.08526481 -0.065240962
5          5           A  0.02675436  0.040652643
6          6           A -0.05021158 -0.014752171
```

By default, for indirect standardization, centers are classified as having low mortality if the (upper bound of the) 50% confidence interval on the standardized risk is smaller than a consensus value. This clinically relevant boundary is by default set at −0.05. Analogously, centers are classified as having high mortality if the (lower bound of the) 50% confidence interval on the standardized risk is larger than the clinically relevant boundary of 0.05. For direct standardization the consensus value is by default 0.8 $\hat{E}(Y)$ for low mortality risks and 1.2 $\hat{E}(Y)$ for high mortality risks. Note that the function `labelCenters()` returns by default 50% confidence intervals in the output, while the function `standardizeRisks()` gives 95% confidence intervals by default.

The clinically relevant boundaries can be adapted by providing specific values for the parameter `lambda` in the labelCenters() function. When lambda is a vector of two elements, the first value determines the consensus value before classifying centers as having low mortality risk, while for high mortality it is the second value. For example, we can implement a consensus value of −0.06 for low mortality risks and 0.02 for high mortality risks as follows:

```
> labeledCenters2 <- labelCenters(standardizedRisks = indirectRisks,
                                  lambda=c(low = -0.06, high = 0.02))
> head(labeledCenters2)

  centerName centerLabel     lowerCI      upperCI
1          1           A -0.03536243 -0.011372117
2          2           A -0.03710506  0.022085690
3          3           A -0.02435179 -0.008265759
4          4           L -0.08526481 -0.065240962
5          5           H  0.02675436  0.040652643
6          6           A -0.05021158 -0.014752171
```

5.2.3 plotRisks()

For indirectly standardized outcomes, the function plotRisks() generates a descriptive scatterplot of the observed against the expected risk under the average care level for patients of that center (Figure 5.2).

```
> plotRisks(standardizedRisks = indirectRisks,
            labeledCenters = labeledCenters)
```

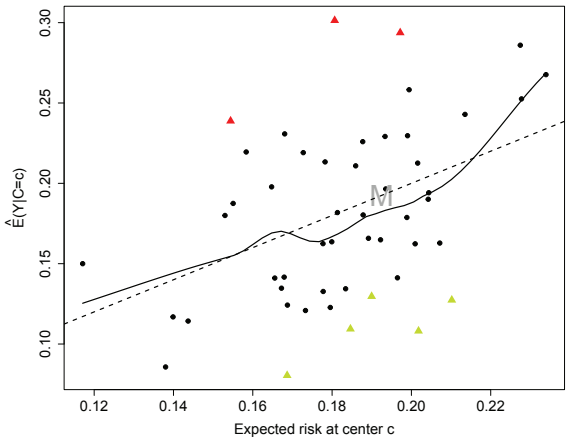


Figure 5.2: Indirect standardization: Plot the observed versus the expected risk under the average care level for patients of that center. The full line represents the best local fit through the points (`loess()` function in R), the dashed line represents the first bisector and 'M' denotes the estimated overall mortality risk $\hat{E}(Y)$.

It illustrates how much the observed risk in each center deviates from the expected risk under the average care level for patients of that center. Large deviation is expected when differences in patient mix are large among centers and is visualised by large deviations from the first bisector. The character 'M' denotes the estimated overall mortality risk $\hat{E}(Y)$.

For directly standardized outcomes, this function generates a descriptive scatterplot of the observed against the estimated directly standardized risk for each center under study.

5.2.4 plotCenterLabels()

Center performance classification can be visualised using the estimated standardized risk and variance per center from the output of `standardizeRisks()` and classification labels from `labelCenters()` (Figure 5.3).

```
> plotCenterLabels(standardizedRisks = indirectRisks,
                   labeledCenters = labeledCenters)
```

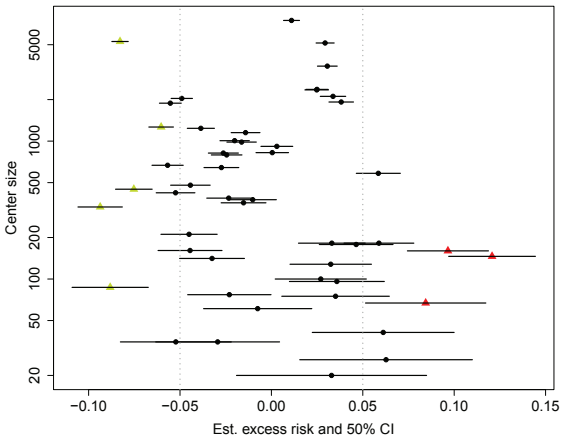


Figure 5.3: Plot the estimated indirectly standardized risk with 50% *confidence limits per center.*

Centers that are classified as having low mortality risk are indicated by green triangles (on the left) and their confidence interval (by default 50%) lies completely below the clinically relevant boundary. Similarly, centers classified with high mortality risk are indicated by red triangles (on the right) and have their confidence interval lying completely above the clinically relevant boundary. Of course, larger centers have narrower estimated confidence intervals than smaller centers.

5.2.5 funnelPlot()

The funnelplot is an internationally recommended plot for comparing institutional performance (Spiegelhalter, 2005a) (Figure 5.4). The estimated standardized risks are plotted against a measure of precision (default is center size). Care centers with an estimated standardized risk lying outside the 95% control limits are flagged as outlying centers. A horizontal line is drawn at the displayed value, for indirect standardization this is at the average of the indirectly standardized excess risks over all centers while for direct standardization this is at the overall mortality risk $\hat{E}(Y)$.

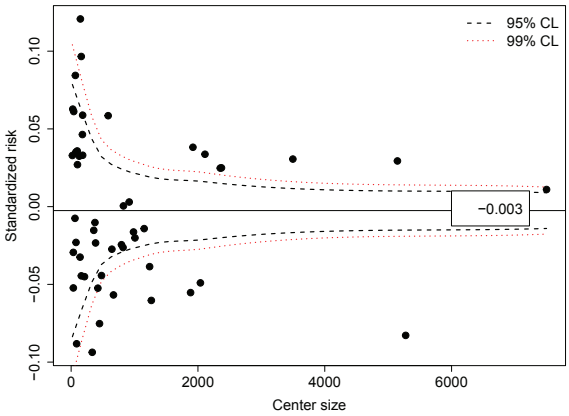


Figure 5.4: Funnelplot for the estimated indirectly standardized risk per center. The horizontal line represents the average over all centers of the indirectly standardized risks.

```
> funnelPlot(standardizedRisks = indirectRisks)
```

CHAPTER 6

Conclusion and Future Research

6.1 Conclusion

Quantifying the quality of hospital care has been fascinating people for several decades (Iezzoni, 2003). Still, the analysis of quality registers faces several statistical challenges, e.g. when some hospitals have a small number of registered patients or when the number and completeness of patient characteristic measurements is limited by the budget. In this thesis, we discussed new statistical methods to handle these limitations and compared their performance with that of current alternatives. Thereby we avoided making ad-hoc adaptations as it may harm confidence in the final reporting and complicate comparisons with earlier study results.

Normal mixed effects models that include fixed effects for patient covariates and random center effects, are commonly used but may lack power to detect small centers with deviating performance due to unintended shrinkage (Normand et al., 1997). In **Chapter 2** we have shown that a Firth corrected fixed effects model only slightly shrinks the fixed center effects towards the overall mean. This approach is thus especially valuable in the presence of small centers

as convergence of the estimation strategy improved over ordinary fixed effects models and more power was retained compared to normal mixed effects models. Weighting observations by the reciprocal of the so-called propensity score, i.e. the probability to be treated in the observed center, may warn the user for undue model extrapolation if patient mix strongly differs across centers. We have investigated the use of an estimator with inverse probability weighting which is doubly robust. It was shown analytically and by simulation that the resulting estimator for the directly standardized risk is indeed doubly robust. So, unbiased estimates of the standardized risks are obtained if either the outcome model or the model for center choice is correctly specified, but not necessarily both (Robins et al., 2007). The model for center choice does not face the issue of extrapolation, so when some centers have only few registered patients the doubly robust estimates may more honestly reflect the uncertainty on the obtained results through wider confidence intervals. Although promising, the doubly robust method was not considered for routine application in its current form due to convergence problems. In summary, we recommended the Firth corrected fixed effects method and used it in all subsequent analysis.

It is known that some centers may perform better on a specific group of patients compared to other centers (Nicholl et al., 2013; Mohammed et al., 2009), for example due to specialized care for the elderly. In **Chapter 3** we found that if this is ignored in the outcome regression model, the directly and indirectly standardized risks are biased when that patient characteristic is very differently distributed between centers, but bias is negligible otherwise. We have also shown that misspecification of the outcome regression model mostly induces bias for the smallest centers for directly standardized risks as opposed to the larger centers for indirectly standardized risks. Given these findings, we can now assess (e.g. via propensity score overlap, which will be explained in Section 6.2) for a clinical register whether common practice of ignoring center-patient interactions is justified and how it will impact the estimated directly or indirectly standardized risks. Justification of including main center effects only, is especially valuable in settings where fitting interactions is simply prohibited by low information content. Still, if including main center effects only is not justified, the question remains how interactions are best included in the outcome model so that the

power to detect deviating performance is preserved and convergence of the estimation strategies is guaranteed.

Of course, high-quality data is most desirable, but in practice it is known to be hindered by budget, personnel or time constraints. Measuring and registering a wide variety of patient characteristics does not only imply a big effort in terms of time and money, but may also damage the quality of registers when more values are missing or incorrect (Shahian et al., 2007). In **Chapter 4** we tackled the problem of selecting the subset of patient covariates that best explains differences in 30-day mortality risks or directly standardized risks and respects a total cost constraint. The stochastic search algorithms allowed for relatively fast and cost-efficient variable selection and could easily handle multiple imputed data sets when some measurements were missing. The flexibility in defining the error function and the options for the search algorithm were a major advantage over other variable selection methods such as the generalized LASSO. These features allow for easy implementation of the stakeholder's preferences and facilitate automated variable selection. Inspection of the selected subset may for example give a clear understanding of which risk factors have a large impact on patient's mortality. One could also decide to restrict future registration to those covariates that resulted in the smallest error on the individual predictions or hospital performance measures. Analytic guidelines were given to evaluate for which proportion of missing values it may be beneficial to include an incompletely measured covariate instead of its surrogate that is less precise but has no missing values. This may either encourage or discourage more complete registration of an expensive covariate depending on whether the expected benefit in error reduction balances the extra registration cost.

We believe that the discussed methodology in this thesis is not only valuable for many clinical registers, but also for school registers or other performance registers. To make the statistical methods available for data analysts in a user-friendly way, we have started building the R package RiskStandard. Given the data, this package estimates, with one click of the mouse, the directly or indirectly standardized risks based on (Firth corrected) logistic outcome regression models that include fixed center and patient-specific effects. Next, the centers with outlying performance can be highlighted, corresponding to the user-defined

values for statistical and clinical significance levels, which depend on whether the user's purpose is public or private reporting. All these results can easily be visualized, using the basic plotting functions documented in **Chapter 5**.

Each of the covered topics were motivated by specific research questions on quantifying quality of care for stroke patients in Sweden (Asplund et al., 2011) or for patients with rectal cancer in a Belgian pilot project (Goetghebeur et al., 2011). The applicability of the developed statistical methods, however, is much wider. On one hand, it is expected that monitoring quality of care based on recorded data will be performed in an increasing range of disease areas. The methods developed here will be directly relevant for that. On the other hand, the interest in monitoring quality of performance is also growing in other fields such as education, where these methods will also allow to evaluate quality in a clearly understandable way and to compare performance between institutions.

6.2 Future Research

6.2.1 Instrumental variable analysis

One of the principal issues when assessing hospital performance is confounding. In this thesis we have studied the use of a multivariate outcome regression model or a doubly robust propensity score model to adjust for confounding (Chapter 2). However, a crucial and unverifiable assumption for these causal inference methods is that the included set of patient covariates is sufficient to adjust for confounding of the center-outcome effect (Hernán and Robins, 2006b). We considered this assumption to be reasonable for Riksstroke because differences in patient mix across hospitals were minimal and a rich set of patient baseline characteristics was measured. Of course, this is certainly not the case for all clinical registers. The plausibility of this assumption both depends on whether the important confounders have been measured and how acute treatment of the disease is. For example, mental illness and cancer are not always treated at the closest center to the patient so that registers for such diseases are expected to show more substantial differences in patient mix than for more acute diseases such as stroke or heart failure. When some important confounders **U** are not

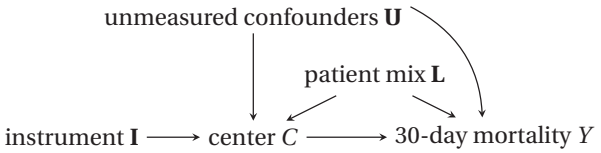


Figure 6.1: A possible instrumental variable setting.

measured (e.g. patient's socio-economic status or geographical pollution levels) the estimated center performance may be biased. In that case, one may consider alternative methods that do not rely on this assumption, such as those using instrumental variables (Hernán and Robins, 2006a).

An instrument is a variable which (i) is, preferably strongly, associated with the center choice and (ii) only affects the 30-day mortality risk via pathways that go through C in Figure 6.1. For example, a vector $\mathbf{I} = (I_1, \ldots, I_m)$ of differential distances to all centers under study has been suggested (Newhouse and McClellan, 1998; Gowrisankaran and Town, 1999), where for a given patient we define I_j as the additional distance to center j, beyond the distance to the nearest center. Naturally, this instrument satisfies the first property as patients are mostly treated at (one of) the nearest hospital(s). The second property, however, assumes that distance to center is not associated with the outcome except through the center choice itself. The latter may be violated if for example some geographical regions have a population with low socio-economic status U which is not measured and clearly associated with a higher mortality risk, while these poor regions do not have many hospitals nearby (extra arrow from **U** to **I** in Figure 6.1). So, the differential distances **I** for patients from poor regions will be much larger than for patients from more wealthy regions. If this is the case, one may still not obtain consistent estimates for the causal center effects (Brookhart et al., 2010; Gowrisankaran and Town, 1999).

In Gowrisankaran and Town (1999) a two-stage linear model for the instrumental variable approach is suggested. In our case it could be postulated as:

$$C_c^* = P(C = c|\mathbf{I}, \mathbf{L}) \quad = \quad \frac{\text{expit}(\boldsymbol{\gamma}_c \mathbf{L} + \boldsymbol{\delta}_c \mathbf{I})}{\sum_{j=1}^{m} \text{expit}(\boldsymbol{\gamma}_j \mathbf{L} + \boldsymbol{\delta}_j \mathbf{I})} \qquad c = 1, \ldots, m \qquad (6.1)$$

$$P(Y = 1|\mathbf{C}^*, \mathbf{L}) \quad = \quad \text{expit}(\boldsymbol{\alpha}\mathbf{L} + \boldsymbol{\beta}\mathbf{C}^*)$$

where $\mathbf{C}^* = (C_1^*, \ldots, C_m^*)$ and approximate estimates for the model parameters can be obtained using two-stage estimation methods (Burgess, 2013). Hospitals are then compared based on the estimated effects $\hat{\boldsymbol{\beta}}$ which are assumed to reflect the given care level. In (6.1) we suggest to fit a separate logistic regression model for each hospital instead of a multinomial model, because we experienced convergence issues when fitting the hospital-specific propensity score models in Chapter 2. Future work may thus investigate how for a binary outcome, the causal center effect is best estimated when a good instrument is available, because two-stage least squares estimation yields only approximate estimates for the model parameters (Vansteelandt et al., 2011). Moreover, it would be interesting to compare both approaches, either relying on the no unmeasured confounders assumption or on the existence of a good instrumental variable. We suggest to make such comparisons based on hospital quality classification because standardized risks cannot easily be obtained using the instrumental variable approach. We also suggest then to study the extent to which there is empirical support for the assumptions given the clinical register and how much results (are expected to) differ between both approaches.

6.2.2 On the methods in this thesis

Doubly robust propensity score methods have been suggested to be promising in **Chapter 2**, especially in settings where patient mix differs substantially across centers or where the outcome regression model may be misspecified. However, small centers resulted in problematically small estimated PS values, especially when its patient case-mix was very different from that of other large centers. Therefore we made an ad-hoc adaptation by stabilizing the vector of propensity scores for each patient by dividing it by the proportion of registered patients at that center. Moreover, for the Riksstroke data convergence issues with multinomial models forced us to build a separate logistic regression model per center. New doubly robust methods, such as the machine learning technique for outcome regression in van der Laan and Gruber (2010) and methods that

are more robust against misspecifaction of both the outcome and PS model, e.g. in Vermeulen and Vansteelandt (2014), show great promise. The construction of confidence intervals with better finite sample performance is an interesting topic for future research.

In **Chapter 3** we have suggested further research on modelling center-patient interactions, when they are necessary. For a small number of hospitals ($m = 4$), a logistic regression model including interaction terms for each hospital and case mix variable in turn, was investigated in Mohammed et al. (2009). When the number of hospitals is large and interactions with many case mix variables together are included, we expect that Firth corrected maximum likelihood estimation may suffer from convergence problems. Even so when the inclusion of interactions induces complete separation, e.g. some hospitals only treated female patients. We believe that a combination of Firth corrected fixed main effects and random interaction effects may reduce the effective model dimension while maintaining power to detect outlying center performance. Thereby it has to be assessed whether assuming a normal distribution for the interaction effects is indeed more plausible than for the main center effects. To our knowledge estimation methods for this combination are not yet investigated. Alternatively, doubly robust models may protect against violation of the assumption of equal covariate effects across centers if the model for center choice is correctly specified. Interactions between center and patient characteristics may also show up due to measurement errors (Nicholl et al., 2013). Blood pressure measurements for example, depend on when and where they are measured. If this is done systematically different in each center, then a patients' blood pressure may be classified as high in one center while in another center it would have been classified as acceptable. Ignoring this, may yield unreliable performance measures.

The stochastic search algorithms in **Chapter 4** focused on minimizing the error function of the patients' individual mortality risk or the directly standardized risk. For the latter, we have discussed that minimizing the error may lack to select confounders of the center-outcome effect. Especially confounders that are strongly related with center choice and weakly with mortality risk may be missed, because covariates were selected based on their predictive value in esti-

mating the mortality risk. Alternative procedures that focus on the selection of confounders are available, such as C-TMLE in van der Laan and Gruber (2010), which does not include a covariate if it blows up the variance on the treatment (i.e. center) effect. On the other hand, in Wilson and Reich (2014) two models are simultaneously penalized: an outcome regression model and a model for the center choice in function of all potential confounders. However, it has not been investigated how covariate costs can be included in these selection procedures.

The R package RiskStandard 0.0.6 implements only a limited number of R functions which in its current version make rather stringent assumptions on the structure of the clinical register, such as covariate values to be missing completely at random. A first step towards a more broadly applicable package would be to make the output for the base function `standardizeRisks()` more generic, so that it can be wrapped in or built on existing R functions. The previously mentioned techniques can then more easily be implemented, such as the stochastic search algorithms for cost-efficient variable selection or multiple imputation of missing covariate values based on the R package MICE (Buuren and Groothuis-Oudshoorn, 2011).

6.2.3 Assessing differences in patient-mix

Throughout this thesis the importance of determining how patient mix differs between centers has been emphasized: to warn the user for unwanted model extrapolation, to confirm the need for modeling center-patient interactions or to identify potential confounders before automatic variable selection. In Chapter 2 we have visually compared the center-specific distribution of each patient characteristic across the centers under study and summarized overlap in patient mix via propensity scores. In Chapter 3 we have also measured the variability of a continuous covariate across centers by the variance of the random intercepts in a random intercept model for $\mathbf{L}$ conditional on center. Still, in practice it is rarely checked how patient mix differs across centers. We believe that this can be encouraged when methods are implemented that allow for easily and clearly doing so, for example as a function in the R package RiskStandard. Especially, multivariate overlap in patient mix should be assessed more routinely.

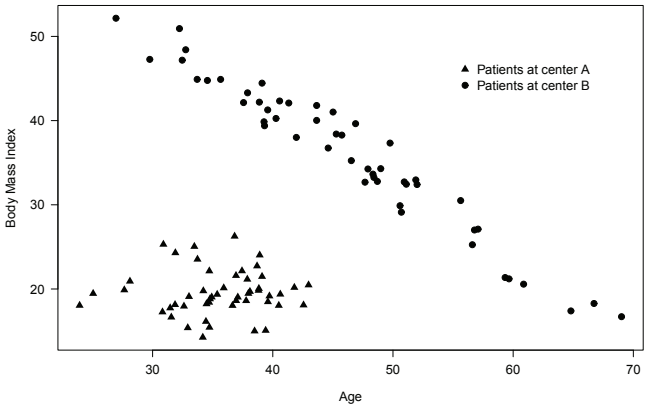


Figure 6.2: The patient's age and body mass index (BMI) for two centers.

Indeed, a univariate overlap does not guarantee that there is also an overlap for combinations of covariates across centers. For example in Figure 6.2, patients at center A and B are clearly not similar when considering age and BMI, although the age distributions for center A and B do overlap, as well as the BMI distributions. In Shahian and Normand (2008) the distribution of propensity scores was used to compare the patient mix for each hospital with that of the pool of patients at the remaining hospitals as well as for selected pairwise comparisons. Although such visual comparisons are very insightful, they may form a subjective criterion and elaborate investigation is needed when there are many hospitals. Future research may focus on developing a statistical test for accordance in patient mix across hospitals.

6.2.4 Longitudinal analysis

We have been monitoring quality of care over a given time period e.g. for Swedish stroke patients treated between 2001 and 2012. Interest may however be in monitoring how the performance of hospitals evolves over time e.g. via annual evaluations (Parry et al., 1998; Campbell et al., 2012). A hospital with deterio-

rating performance over time should evidently receive more attention than a hospital with ameliorating performance. Most of the developed methods can still be applied on an annual subset of the data, although this certainly has consequences. First, the sample size will drop, and especially for small hospitals the size may become problematically small. If the Firth corrected maximum likelihood estimation does not converge, a separate normal random effects distribution per group of centers (e.g. urban/rural hospitals, university/non-university hospitals) or another distribution than the normal distribution for the random center effects may provide an alternative (Ash et al., 2012). A second concern is that measures of hospital performance should not blindly be compared between different time periods: It has to be kept in mind that the reference population may be shifted. For a given hospital, estimated differences in DSRs or ISRs over years may be explained by a shift in the overall population under study for DSRs or the center-specific population for ISRs. A more stable reference for DSRs could for example be obtained by taking a 5-year window of treated patients as reference population. More interesting however would be to make direct comparisons of center performance between several years. This requires taking into account the correlation between performance measures on successive time points. For example, given 4 successive time points, a continuous outcome Y_{ij} for a patient i at time point j can be modeled as:

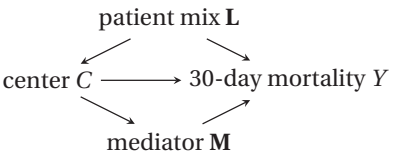
$$\begin{aligned} Y_{ij} &= \boldsymbol{\beta}\mathbf{L}_i + \sum_{c=1}^{m} \psi_c I(C_i = c) + \varepsilon_{ij} \qquad i = 1,\ldots,n; j = 1,\ldots,4 \qquad (6.2)\\ \psi_c &\sim N(\mu, \sigma^2_\psi) \qquad c = 1,\ldots,m \\ \varepsilon_i &\sim N(\mathbf{0}, \sigma^2_\varepsilon \mathbf{V}) \qquad i = 1,\ldots,n \\ \mathbf{V} &= \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}, \end{aligned}$$

where $(\psi_1,\ldots,\psi_m)$ and $\varepsilon_i = (\varepsilon_{i1},\ldots,\varepsilon_{i4})$ are independent and a first-order autoregressive structure for the variance matrix $\mathbf{V}$ is assumed (Fitzmaurice et al., 2004). Pairwise comparisons between two years may then help to examine how

the correlation actually decreases when the period between two time points increases.

6.2.5 Mediation analysis

Identifying the reasons for different care levels between hospitals plays an important role in improving quality of care. Some differences may be explained by hospital characteristics, such as its size or medical infrastructure (Goldstein et al., 2002; Saposnik et al., 2007; Tung et al., 2015). In a causal inference framework these hospital attributes can be seen as mediators **M** of the center-outcome effect.



Including the hospital attributes in the logistic regression model can be based on advanced methodology for mediation analysis (VanderWeele and Vansteelandt, 2010). The natural direct and indirect effect can distinguish between the effect of the given care level and the effect of the mediator on the mortality risk, but much less is known on how the measures for hospital performance are then best defined. For example, for the indirectly standardized risks for a given center, the mediator value could be set at the value that would be observed under the care level for that center. The ISR for center c would then compare the observed and expected risk while the mediator value is as observed in center c.

To conclude, in this thesis we have provided some statistical tools to guide the analysis of clinical registers. We have shown how and when they may facilitate quantifying the quality of care by analyzing the Riksstroke data. However, it is clear that there is still much more to learn about this interesting topic.

CHAPTER 7

Samenvatting

In deze thesis onderzoeken we hoe de analyse van uitkomsten, zoals overleving van patiënten met een beroerte, kan helpen om de kwaliteit van de zorg te vergelijken tussen ziekenhuizen. Naar analogie met de buurlanden, groeit de vraag naar kwaliteitscontrole in ziekenhuizen maar ook bijvoorbeeld in woonzorgcentra en scholen, en dit zowel vanuit de overheid en de patiënten als de centra zelf. Gezien de mogelijks grote impact van gerapporteerde resultaten vraagt dit om een zorgvuldige statistische analyse van de beschikbare data, zoals besproken in **Hoofdstuk 1**. Om het oorzakelijk (i.e. causaal) effect van de zorgkwaliteit in een ziekenhuis op de gekozen uitkomst te kunnen schatten, moet er gecorrigeerd worden voor patiëntkenmerken, zoals leeftijd en ernst van de ziekte. Dit is nodig omdat ze de uitkomst kunnen beïnvloeden en mogelijks verschillend verdeeld zijn over de ziekenhuizen. Zonder die correctie kan een ziekenhuis dat vooral oudere patiënten behandelt een hoger sterftecijfer vertonen, ook al is de zorgkwaliteit er uitstekend. De onderzoeksvragen in deze thesis zijn voornamelijk geïnspireerd vanuit de analyse van het Zweedse kwaliteitsregister voor acute beroertes, Riksstroke (`http://www.riksstroke.org/eng/`), maar de ontwikkelde methodes zijn uiteraard meer algemeen toepasbaar. Om rekening

te houden met patiëntkenmerken zullen, afhankelijk van de onderzoeksvraag, direct of indirect gestandaardiseerde risico's gebruikt worden als maat voor de performantie.

Het is bewezen dat het schatten van gestandaardiseerde risico's met behulp van het populaire normal mixed effects model, de geschatte kwaliteit van ziekenhuizen richting het gemiddelde kan trekken, waardoor afwijkende performantie vaak niet gedetecteerd wordt (Normand et al., 1997; Ash et al., 2012). In **Hoofdstuk 2** onderzochten we daarom het gebruik van een Firth gecorrigeerd fixed effects model en vonden daarbij slechts lichte krimping van de centrumeffecten richting het globale gemiddelde. Deze aanpak is dus bijzonder nuttig wanneer sommige ziekenhuizen een klein aantal patiënten hebben geregistreerd, aangezien de convergentie van de schattingsstrategie beter is dan voor fixed effects modellen en er een betere detectie van afwijkende performantie is dan voor normal mixed effects modellen. Een tweede aspect dat we behandelen is onbewuste model-extrapolatie bij het schatten van bijvoorbeeld direct gestandaardiseerde risico's, vooral wanneer de patiëntenmix sterk verschilt tussen ziekenhuizen. Extrapolatie in combinatie met het gebruik van foute statistische modellen kan vertekende resultaten opleveren met een onderschatte onzekerheid. Daarom onderzochten we een methode die observaties weegt met het omgekeerde van de zogenaamde propensity score, d.w.z. de kans om in het geobserveerde ziekenhuis behandeld te worden (Shahian and Normand, 2008). De onderzochte dubbel robuuste methode is meer beschermd tegen het gebruik van foute modellen (Robins et al., 2007) en zal, als de propensity score zeer klein is, de gebruiker waarschuwen voor extrapolatie via opgeblazen variantieschattingen. Hoewel veelbelovend, raden we op basis van de bekomen resultaten de Firth gecorrigeerde fixed effects methode aan.

Gemeenschappelijke correcties voor verschillen in patiëntenmix veronderstellen doorgaans dat het effect van zorgniveau op de uitkomst constant is over patiëntkenmerken (Ohlssen et al., 2007b; Shahian and Normand, 2008). In de praktijk is dit echter niet altijd het geval, bijvoorbeeld door gespecialiseerde zorg voor ouderen (Nicholl et al., 2013; Mohammed et al., 2009). Als er in dat geval geen interacties tussen centrum en patiënt in het uitkomst regressiemodel worden opgenomen, dan vonden we in **Hoofdstuk 3** dat de direct en indirect

gestandaardiseerde risico's enkel vertekend zijn indien de verdeling van het betreffende patiëntkenmerk sterk verschilt over de centra, anders is de vertekening verwaarloosbaar. Het kunnen rechtvaardigen van de gangbare praktijk is vooral belangrijk in situaties waarbij het simpelweg onmogelijk is om deze interacties in het model te schatten, omdat er onvoldoende informatie beschikbaar is in kleine ziekenhuizen, zie bijvoorbeeld Ash et al. (2012).

In **Hoofdstuk 4** onderzochten we ook hoe het aantal (dure, genetische) metingen - en dus de kost per patiënt - kan beperkt worden wanneer we de uitkomst willen voorspellen voor individuele patiënten of voor het gestandaardiseerde risico bij het meten van ziekenhuiskwaliteit. Stochastische zoekalgoritmes laten toe om een relatief snelle en kost-efficiënte variabelenselectie uit te voeren en ze kunnen gemakkelijk overweg met meervoudig geïmputeerde datasets wanneer sommige metingen ontbreken. We hebben bovendien geïllustreerd hoe de rekentijd verder gereduceerd kan worden door voorafgaand een kost-gebaseerd generalized LASSO algoritme uit te voeren.

Omdat we geloven in de brede toepasbaarheid van de statistische methodes in deze thesis, stellen we ze beschikbaar via het R-pakket RiskStandard (`www.cvstat.ugent.be`), zoals gedocumenteerd in **Hoofdstuk 5**.

CHAPTER 8

Summary

In this thesis, we examine how the analysis of quality outcomes, such as 30-day mortality for patients with acute stroke, can help compare the quality of care between hospitals. As in the neighboring countries, the demand for quality control in hospitals is growing but also, for example for residential care centers and schools, both by government and patients as well as centers themselves. Given the potentially large impact of reported results, this requires a careful statistical analysis of the available data, as discussed in **Chapter 1**. To estimate the causal effect of the quality of care on the outcome of interest, we have to control for differences between patients on admission, such as age and initial disease severity. This is necessary because they may influence the outcome and they are possibly distributed differently across centers. Otherwise, hospitals treating mostly elderly patients may show higher mortality risks, even though the given care is excellent. The research questions in this thesis were mostly inspired by the analysis of the Swedish register for acute stroke care, Riksstroke (`http://www.riksstroke.org/eng/`), but the discussed methods are more generally applicable. To account for measured patient characteristics we will use, depending on the research question, directly or indirectly standard-

ized risks as performance measure.

It has been proven that when standardized risks are estimated based on the popular normal mixed effects model, the estimated quality of care may be shrunken towards the average, often masking outlying performance of hospitals (Normand et al., 1997; Ash et al., 2012). In **Chapter 2** we therefore investigated the use of a Firth corrected fixed effects model and found little shrinkage of the center effects towards the overall mean. This approach is thus particularly valuable when some centers have a small number of registered patients since the convergence of this estimation strategy is better than for fixed effects models and a better detection of outlying performance is obtained than for normal mixed effects models. Secondly, we investigate undue model extrapolation when estimating for example, directly standardized risks, especially if patient mix differs substantially between hospitals. Extrapolation in combination with the use of misspecified statistical models can yield biased results with an underestimated uncertainty. Therefore, we examined a method that weights observations by the inverse of the so-called propensity score, i.e. the probability to be treated in the observed center (Shahian and Normand, 2008). The investigated doubly robust method is protected against model misspecification (Robins et al., 2007) and, if the propensity score is very small, the user will be warned for extrapolation via inflated variance estimates. Although promising, the obtained results suggested to use the Firth corrected fixed effects method.

Common adjustments for differences in patient mix generally assume that the effect of the given care level on the outcome is constant across patient groups (Ohlssen et al., 2007b; Shahian and Normand, 2008). In practice, however, this may be violated when some centers are for example specialized in care for the elderly (Nicholl et al., 2013; Mohammed et al., 2009). If then no center-patient interactions are included in the outcome regression model, we found in **Chapter 3** that the directly and indirectly standardized risks will only be biased if the distribution of that patient characteristic differs substantially across centers, otherwise bias is negligible. Being able to justify common practice is especially important in settings where it is simply impossible to estimate these interactions in the model, because insufficient information is available in small hospitals, for example see Ash et al. (2012).

In **Chapter 4** we also examined how the number of (expensive, genetic) measurements - and thus the cost per patient - can be reduced when predicting individual patient outcomes or estimating standardized risks for hospital quality evaluation. Stochastic search algorithms allow for a relatively quick and cost-efficient variable selection and they can easily handle multiple imputed datasets when some measurements are missing. We have also illustrated how the search time can be further reduced by a priori performing a cost-efficient generalized LASSO search.

Because we believe in the broad applicability of the statistical methods in this thesis, we have made them available via the R-package RiskStandard (`www.cvstat.ugent.be`), as documented in **Chapter 5**.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis (Second Edition)*. Wiley, New York.

Ash, A. S., Fienberg, E., Louis, A., Norm, S.-l. T., Stukel, A., and Utts, P. J. (2012). Statistical issues in assessing hospital performance. The COPSS-CMS White Paper Committee. *Citeseer*.

Asplund, K., Hulter Asberg, K., Appelros, P., Bjarne, D., Eriksson, M., and Johansson, A. e. a. (2011). The Riks-Stroke story: building a sustainable national register for quality assessment of stroke care. *International Journal of Stroke*, 6: 6–99.

Austin, P., Alter, D., and Tu, J. (2003). The use of fixed- and random-effects models for classifying hospitals as mortality outliers: A Monte Carlo assessment. *Medical Decision Making*, 23(6): 526–539.

Black, N. (2010). Assessing the quality of hospitals. *BMJ*, 340.

Bos, V., Kunst, A. E., Garssen, J., and Mackenbach, J. P. (2005). Socioeconomic inequalities in mortality within ethnic groups in the Netherlands, 1995–2000. *Journal of epidemiology and community health*, 59(4): 329–335.

Brookhart, M. A., Stürmer, T., Glynn, R. J., Rassen, J., and Schneeweiss, S. (2010). Confounding control in healthcare database research: challenges and potential approaches. *Medical care*, 48(6 0): S114.

Brookhart, M. A. and van der Laan, M. J. (2006). A semiparametric model selection criterion with applications to the marginal structural model. *Computational statistics & data analysis*, 50(2): 475–498.

Burgess, S. (2013). Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Statistics in Medicine*, 32(27): 4726–4747.

Buuren, S. and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45(3).

Campbell, M., Jacques, R., Fotheringham, J., Maheswaran, R., and Nicholl, J. (2012). Developing a summary hospital mortality index: retrospective analysis in English hospitals over five years. *British Medical Journal*, 344(e1001).

Campbell, S. M., Roland, M. O., and Buetow, S. A. (2000). Defining quality of care. *Social science & medicine*, 51(11): 1611–1625.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Clinical Indicators Team (2015). Indicator specification: Summary Hospital-level Mortality Indicator. http://www.hscic.gov.uk. Accessed: 2015-09-22.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

De Beukelaer, H., Davenport, G. F., De Meyer, G., and Fack, V. (2015). JAMES: A modern object-oriented Java framework for discrete optimization using local search metaheuristics. *4th International Symposium And 26th National Conference On Operational Research. Chania: Hellenic Operational Research Society*, pages 134–138.

DeLong, E., Peterson, E., DeLong, D., Muhlbaier, L., Hackett, S., and Mark, D. (1997). Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine*, 16(23): 2645–2664.

Firth, D. (1992). Bias reduction, the Jeffreys prior and GLIM. In Fahrmeir, E. L., Francis, B., Gilchrist, R., and Tutz, G., editors, *Advances in GLIM and Statistical Modelling*, pages 91–100. New York: Springer-Verlag.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1): 27–38.

Fitzmaurice, G., Laird, N., and Ware, J. (2004). *Applied Longitudinal Analysis*. John Wiley & Sons, USA.

Fouskakis, D. and Draper, D. (2008). Comparing stochastic optimization methods for variable selection in binary outcome prediction, with application to health policy. *Journal of the American Statistical Association*, 103(484): 1367–1381.

Fouskakis, D., Ntzoufras, I., and Draper, D. (2009). Population-based reversible jump Markov chain Monte Carlo methods for Bayesian variable selection and evaluation under cost limit restrictions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(3): 383–403.

Freeman, T. (2002). Using performance indicators to improve health care quality in the public sector: a review of the literature. *Health Services Management Research*, 15(2): 126–137.

Gatsonis, C., Normand, S.-L., Liu, C., and Morris, C. (1993). Geographic variation of procedure utilization: a hierarchical model approach. *Medical Care*, 31(5): YS54–YS59.

Gatsonis, C. A., Epstein, A. M., Newhouse, J. P., Normand, S.-L., and McNeil, B. J. (1995). Variations in the utilization of coronary angiography for elderly patients with an acute myocardial infarction: an analysis using hierarchical logistic regression. *Medical Care*, 33(6): 625–642.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423): 881–889.

Glover, F. and Kochenberger, G. A. (2010). *Handbook of metaheuristics*. Springer Science & Business Media.

Goetghebeur, E., Van Rossem, R., Baert, K., Vanhoutte, K., Boterberg, T., Demetter, P., De Ridder, M., Harrington, D., Peeters, M., Storme, G., Verhulst, J., Vlayen, J., Vrijens, F., Vansteelandt, S., and Ceelen, W. (2011). Quality insurance of rectal cancer - phase 3: statistical methods to benchmark centers on a set of quality indicators, Good Clinical Practice (GCP). *Belgian Health Care Knowledge Centre (KCE)*, KCE Report 161C, D/2011/10.273/40: 1–142.

Goldstein, S. M., Ward, P. T., Leong, G. K., and Butler, T. W. (2002). The effect of location, strategy, and operations technology on hospital performance. *Journal of Operations Management*, 20(1): 63–75.

Gowrisankaran, G. and Town, R. J. (1999). Estimating the quality of care in hospitals using instrumental variables. *Journal of health economics*, 18(6): 747–767.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4): 711–732.

Greenland, S. (2008). Invited commentary. *American Journal of Epidemiology*, 167: 523–529.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1): 29–46.

Gross, P. A., Greenfield, S., Cretin, S., Ferguson, J., Grimshaw, J., Grol, R., Klazinga, N., Lorenz, W., Meyer, G. S., Riccobono, C., Schoenbaum, S. C., Schyve, P., and Shaw, C. (2001). Optimal methods for guideline implementation: conclusions from Leeds Castle meeting. *Medical care*, 39(8): II–85.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2): 481–488.

He, K., Kalbfleisch, J. D., Li, Y., and Li, Y. (2013). Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. *Lifetime data analysis*, 19(4): 490–512.

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, 25(24): 4216–4226.

Hernán, M. and Robins, J. (2006a). Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4): 360–372.

Hernán, M. A. and Robins, J. M. (2006b). Estimating causal effects from epidemiological data. *Journal of epidemiology and community health*, 60(7): 578–586.

Hocking, R. R. (1976). A Biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, pages 1–49.

Iezzoni, L. I. (1997). Assessing quality using administrative data. *Annals of internal medicine*, 127(8 Pt 2): 666–674.

Iezzoni, L. I. (2003). *Risk adjustment for measuring health care outcomes*. Chicago: Health Administration Press, 2003.

Jha, A. K., Li, Z., Orav, E. J., and Epstein, A. M. (2005). Care in US hospitals – the Hospital Quality Alliance program. *New England Journal of Medicine*, 353(3): 265–274.

Kahn, K. L., Rubenstein, L. V., Draper, D., Kosecoff, J., Rogers, W. H., Keeler, E. B., and Brook, R. H. (1990). The effects of the DRG-based prospective payment system on quality of care for hospitalized medicare patients: an introduction to the series. *The Journal of the American Medical Association*, 264(15): 1953–1955.

Kalbfleisch, J. D. and Wolfe, R. A. (2013). On monitoring outcomes of medical providers. *Statistics in Biosciences*, 5(2): 286–302.

Keeler, E. B., Rubenstein, L. V., Kahn, K. L., Draper, D., Harrison, E. R., McGinty, M. J., Rogers, W. H., and Brook, R. H. (1992). Hospital characteristics and quality of care. *The Journal of the American Medical Association*, 268(13): 1709–1714.

Keiding, N. and Clayton, D. (2014). Standardization and control for confounding in observational studies: a historical perspective. *Statistical Science*, 29(4): 529–558.

Kessels, R., Jones, B., and Goos, P. (2013). An argument for preferring Firth bias-adjusted estimates in aggregate and individual-level discrete choice modeling. Technical Report D/2013/1169/013, University of Antwerp.

Knol, M. J., Janssen, K. J., Donders, A. R. T., Egberts, A. C., Heerdink, E. R., Grobbee, D. E., Moons, K. G., and Geerlings, M. I. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of clinical epidemiology*, 63(7): 728–736.

Kosmidis, I. (2011). brglm: Bias reduction in generalized linear models.

Kosmidis, I. and Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika*, 96(4): 793–804.

Kressner, M., Bohe, M., and Cedermark, B. e. a. (2009). The impact of hospital volume on surgical outcome in patients with rectal cancer. *Diseases ot the Colon & Rectum*, 52(9): 1542–1549.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5): 1–26.

Leckie, G. and Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society*, 172(4): 835–851.

Lilford, R., Mohammed, M. A., Spiegelhalter, D., and Thomson, R. (2004). Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *The Lancet*, 363(9415): 1147–1154.

Lilford, R. and Pronovost, P. (2010). Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. *BMJ*, 340.

Lin, C. B., Peterson, E. D., Smith, E. E., Saver, J. L., Liang, L., Xian, Y., Olson, D. M., Shah, B. R., Hernandez, A. F., Schwamm, L. H., and Fonarow, G. C. (2012). Emergency medical service hospital prenotification is associated with improved evaluation and treatment of acute ischemic stroke. *Circulation: Cardiovascular Quality and Outcomes*, 5(4): 514–522.

Lindmark, A., Glader, E.-L., Asplund, K., Norrving, B., and Eriksson, M. (2014). Socioeconomic disparities in stroke case fatality–Observations from Riks-Stroke, the Swedish stroke register. *International Journal of Stroke*, 9(4): 429–436.

Little, R. J. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*, 87(420): 1227–1237.

Liu, J. and Gustafson, P. (2008). On average predictive comparisons and interactions. *International Statistical Review*, 76(3): 419–432.

Manktelow, B. N., Evans, T. A., and Draper, E. S. (2014). Differences in case-mix can influence the comparison of standardised mortality ratios even with optimal risk adjustment: an analysis of data from paediatric intensive care. *BMJ quality & safety*.

Mannion, R. and Goddard, M. (2003). Public disclosure of comparative clinical performance data: lessons from the Scottish experience. *Journal of evaluation in clinical practice*, 9(2): 277–286.

Mant, J. (2001). Process versus outcome indicators in the assessment of quality of health care. *International Journal for Quality in Health Care*, 13(6): 475–480.

Mehta, R. H., Peterson, E. D., and Califf, R. M. (2007). Performance measures have a major effect on cardiovascular outcomes: a review. *The American journal of medicine*, 120(5): 398–402.

Moerkerke, B. and Goetghebeur, E. (2006). Selecting 'significant' differentially expressed genes from the combined perspective of the null and the alternative. *Journal of Computational Biology*, 13(9): 1513–1531.

Mohammed, M., Manktelow, B., and Hofer, T. (2012). Comparison of four methods for deriving hospital standardised mortality ratios from a single hierarchical logistic regression model. *Statistical Methods in Medical Research*, page epub ahead of print.

Mohammed, M. A., Deeks, J. J., Girling, A., Rudge, G., Carmalt, M., Stevens, A. J., and Lilford, R. J. (2009). Evidence of methodological bias in hospital standardised mortality ratios: retrospective database study of English hospitals. *BMJ*, 338: b780.

Musoro, J. Z., Zwinderman, A. H., Puhan, M. A., ter Riet, G., and Geskus, R. B. (2014). Validation of prediction models based on lasso regression with multiply imputed data. *BMC medical research methodology*, 14(1): 116.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied linear statistical models*, volume 4. Irwin Chicago.

Newhouse, J. P. and McClellan, M. (1998). Econometrics in outcomes research: the use of instrumental variables. *Annual review of public health*, 19(1): 17–34.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, 16(1): 1–32.

Nicholl, J., Jacques, R. M., and Campbell, M. J. (2013). Direct risk standardisation: a new method for comparing casemix adjusted event rates using complex models. *BMC medical research methodology*, 13(133).

Normand, S., Glickman, M., and Gatsonis, C. (1997). Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association*, 92(439): 803–814.

Normand, S.-L. and Shahian, D. M. (2007). Statistical and clinical aspects of hospital outcomes profiling. *Statistical Science*, 22(2): 206–226.

O'Brien, S. and Dunson, D. (2004). Bayesian multivariate logistic regression. *Biometrics*, 60: 739–746.

Ohlssen, D., Sharples, L., and Spiegelhalter, D. (2007a). Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Statistics in Medicine*, 26: 2088–2112.

Ohlssen, D., Sharples, L., and Spiegelhalter, D. (2007b). A hierarchical modelling framework for identifying unusual performance in health care providers. *Journal of Royal Statistical Society A*, 170(4): 865–890.

Parry, G., Gould, C., McCabe, C., and Tarnow-Mordi, W. (1998). Annual league tables of mortality in neonatal intensive care units: longitudinal study. International Neonatal Network and the Scottish Neonatal Consultants and Nurses Collaborative Study Group. *BMJ*, 316: 1931–1935.

Peduzzi, P., Concato, J., Kemper, E., Holford, T., and Feinstein, A. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12): 1373–1379.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, pages 20–22.

Roalfe, A. K., Holder, R. L., and Wilson, S. (2008). Standardisation of rates using logistic regression: a comparison with the direct method. *BMC health services research*, 8(1): 1–275.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, 22(4): 544–559.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1): 15–32.

Rubin, D. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127(8S): 757–763.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3): 581–592.

Saposnik, G., Baibergenova, A., O'Donnell, M., Hill, M., Kapral, M., Hachinski, V., and behalf of the Stroke Outcome Research Canada (SORCan) Working Group, O. (2007). Hospital volume and stroke outcome: Does it matter? *Neurology*, 69(11): 1142–1151.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, 8(1): 3–15.

Shahian, D. and Normand, S. (2008). Comparison of "risk-adjusted" hospital outcomes. *Circulation: Journal of the American Heart Association*, 117: 1955–1963.

Shahian, D. M., Blackstone, E. H., Edwards, F. H., Grover, F. L., Grunkemeier, G. L., Naftel, D. C., Nashef, S. A., Nugent, W. C., and Peterson, E. D. (2004). Cardiac surgery risk models: a position article. *The Annals of Thoracic Surgery*, 78(5): 1868–1877.

Shahian, D. M., Normand, S.-L., Torchiana, D. F., Lewis, S. M., Pastore, J. O., Kuntz, R. E., and Dreyer, P. I. (2001). Cardiac surgery report cards: comprehensive review and statistical critique. *The Annals of Thoracic Surgery*, 72(6): 2155–2168.

Shahian, D. M., Silverstein, T., Lovett, A. F., Wolf, R. E., and Normand, S.-L. T. (2007). Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards. *Circulation*, 115(12): 1518–1527.

Shaw, J., Taylor, R., and Dix, K. (2015). Uses & abuses of performance data in healthcare. *Dr Foster*, pages 1–41.

Silber, J. H., Rosenbaum, P. R., Brachet, T. J., Ross, R. N., Bressler, L. J., Even-Shoshan, O., Lorch, S. A., and Volpp, K. G. (2010). The hospital compare mor-

tality model and the volume–outcome relationship. *Health services research*, 45(5p1): 1148–1167.

Spahn, M., Boxler, S., Joniau, S., Moschini, M., Tombal, B., and Karnes, R. J. (2015). What is the need for prostatic biomarkers in prostate cancer management? *Current urology reports*, 16(10): 1–7.

Spiegelhalter, D. (2005a). Funnel plots for comparing institutional performance. *Statistics in Medicine*, 24(8): 1185–1202.

Spiegelhalter, D. (2005b). Handling over-dispersion of performance indicators. *Quality and Safety in Health Care*, 14(5): 347–351.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338: b2393.

Steyerberg, E. W., Harrell, F. E., Borsboom, G. J., Eijkemans, M., Vergouwe, Y., and Habbema, J. D. F. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology*, 54(8): 774–781.

Stroke Board Team, R. (2011). Brief summary of data collected in 2011. Accessed: 2015-10-19.

The ATLANTIS, ECASS, and NINDS rt-PA Study Group Investigators (2004). Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. *Lancet*, 363(9411): 768–774.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *Ann. Statist.*, 39(3): 1335–1371.

Trusheim, M. R., Berndt, E. R., and Douglas, F. L. (2007). Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nature Reviews Drug Discovery*, 6(4): 287–293.

Tung, Y.-C., Jeng, J.-S., Chang, G.-M., and Chung, K.-P. (2015). Processes and outcomes of ischemic stroke care: the influence of hospital level of care. *International Journal for Quality in Health Care*, 27(4): 260–266.

van der Heijden, G. J., Donders, A. R. T., Stijnen, T., and Moons, K. G. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of clinical epidemiology*, 59(10): 1102–1109.

van der Laan, M. and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics*, 6(1).

VanderWeele, T. J., Mukherjee, B., and Chen, J. (2012). Sensitivity analysis for interactions under unmeasured confounding. *Statistics in Medicine*, 31(22): 2552–2564.

VanderWeele, T. J. and Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American journal of epidemiology*, 172(12): 1339–1348.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2010). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1): 7–30.

Vansteelandt, S., Bowden, J., Babanezhad, M., and Goetghebeur, E. (2011). On instrumental variables estimation of causal odds ratios. *Statistical Science*, 26(3): 403–422.

Varewyck, M., Goetghebeur, E., Eriksson, M., and Vansteelandt, S. (2014). On shrinkage and model extrapolation in the evaluation of clinical center performance. *Biostatistics*, 15(4): 651–664.

Varewyck, M., Vansteelandt, S., Eriksson, M., and Goetghebeur, E. (2015). On the practice of ignoring center-patient interactions in evaluating hospital performance. *Statistics in Medicine*.

Vergouwe, Y., Royston, P., Moons, K. G., and Altman, D. G. (2010). Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of clinical epidemiology*, 63(2): 205–214.

Vermeulen, K. and Vansteelandt, S. (2014). Biased-reduced doubly robust estimation. *Journal of the American Statistical Association*, page epub ahead of print.

Wang, C., Parmigiani, G., and Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68(3): 661–671.

White, I. and Carlin, J. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29(28): 2920–2931.

Wilson, A. and Reich, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics*, 70(4): 852–861.

Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27(17): 3227–3246.

World Health Organization and others (2003). Quality and accreditation in health care services: a global review. *Geneva: World Health Organization*.